

# Learning Discriminative Bayesian Networks from High-dimensional Continuous Neuroimaging Data

Luping Zhou, Lei Wang, Lingqiao Liu, Philip Ogunbona, Dinggang Shen

**Abstract**—Due to its causal semantics, Bayesian networks (BN) have been widely employed to discover the underlying data relationship in exploratory studies, such as brain research. Despite its success in modeling the probability distribution of variables, BN is naturally a generative model, which is not necessarily discriminative. This may cause the ignorance of subtle but critical network changes that are of investigation values across populations. In this paper, we propose to improve the discriminative power of BN models for continuous variables from two different perspectives. This brings two general discriminative learning frameworks for Gaussian Bayesian networks (GBN). In the first framework, we employ Fisher kernel to bridge the generative models of GBN and the discriminative classifiers of SVMs, and convert the GBN parameter learning to Fisher kernel learning via minimizing a generalization error bound of SVMs. In the second framework, we employ the max-margin criterion and build it directly upon GBN models to explicitly optimize the classification performance of the GBNs. The advantages and disadvantages of the two frameworks are discussed and experimentally compared. Both of them demonstrate strong power in learning discriminative parameters of GBNs for neuroimaging based brain network analysis, as well as maintaining reasonable representation capacity. The contributions of this paper also include a new Directed Acyclic Graph (DAG) constraint with theoretical guarantee to ensure the graph validity of GBN.

**Index Terms**—Bayesian network, discriminative learning, Fisher kernel learning, max-margin, brain network.

## 1 INTRODUCTION

As an important probabilistic graphical model, Bayesian network (BN) has been used to model the probability distribution of a set of random variables for a wide spectrum of applications, e.g., diagnosis, troubleshooting, web mining, meteorology and bioinformatics. It combines graph representation with Bayesian analysis, providing an effective way to model and infer the conditional dependency of the variables. A BN has to be a directed acyclic graph (DAG). Two factors characterize a BN, i.e., the structure of the network (the presence / absence of edges in the graph) and the parameters of the probability distribution. Recent research of BN focuses on how to learn the structure and the parameters of BN directly from the data.

The approaches of learning BN structures can be roughly categorized into the constraint-based, the score-based, and the hybrid approaches. The constraint-based approaches use a series of conditional independence testing to ensure the model structure is consistent with the conditional independency entailed by the observations. Methods in this class include the IC algorithm [1], PC algorithm [2], and more recent methods [3], [4]. Score-based approaches define a scoring function over the

space of candidate DAGs and optimize this function through certain search strategies. Methods in this class vary with scoring criteria, e.g., the posterior probability [5], [6], [7] and the minimum description length [8], or vary with search strategies, e.g., the heuristic search [9] and the Monte Carlo methods [5]. Hybrid approaches usually employ constraint-based methods to prune the search space of DAG structures and consequently restrict a subsequent score-based search [10], [11]. Many existing BN learning methods, such as LIMB-DAG [12], MMHC [10], TC and TC-bw [13], comprise of two stages: the identification of candidate parent sets in the first stage and the further pruning of them based on certain criteria in the second stage. Despite the mitigation of computational complexity, a drawback arises that if a parent node is missed in the first stage, it will never be recovered in the second stage [14]. To address this issue, one-stage learning process has been preferred in recent research work [14], [15]. In these studies, based on Gaussian Bayesian network (GBN), the parent sets of all variables are learned together to optimize a LASSO-based score function in a single stage. The related optimization problems are solved either approximately [14] or exactly [15]. They have demonstrated improved reliability of BN edge identification over traditional two stage methods.

Although BN is naturally a generative method, it has also been used in classification tasks for diagnostic or predictive purposes. A straightforward usage is to train each class a BN and classify a new sample into the class with the highest likelihood value [14]. Another kind of

• L. Zhou, L. Wang and P. Ogunbona are with the School of Computing and Information Technology, University of Wollongong, NSW 2500, Australia. E-mail: lupingz, leiw, philipo@uow.edu.au. L. Liu is with Shool of Computer Science, University of Adelaide, Australia. D. Shen is with Department of Radiology, University of North Carolina at Chapel Hill, USA.

approaches trains “Bayesian network classifiers” with discriminative objective functions [16], [17], [18]. In these approaches, usually a single BN is learned to optimize the discrimination performance. Either the structure or the parameters of the BN are adjusted to reflect the class difference for better classification. Therefore, the resulting BN does not model the distribution of any individual class. The “Bayesian network classifiers” in [16], [17], [18] are designed for discrete variables of multinomial distribution. They still inherit the two-stage learning process, i.e., have to predefine candidate parent sets as mentioned above.

Learning BN from the data faces new challenges in exploratory domains, such as brain research, where the mechanism of brain and mental diseases remain unclear and need to be explored. These domains usually cater for both interpretation and discrimination. “Interpretation” requires interpretable models of the data and the findings explained by domain language rather than mathematical terms. This requirement comes from the demand of understanding the domain problems. “Discrimination” requires the models to have sufficient discriminative power to distinguish groups of interest (such as identifying the diseased from the healthy), for the purpose of prediction. To some extent, a high accuracy of the predictive model also provides a measure of the amount of information captured by that model.

Being a generative method, BN represents the distribution of the data and is naturally amenable for interpretation. However, it is known that generative methods are not necessarily discriminative. They are prone to emphasizing major structures that are shared within each group, and neglecting the subtle but critical changes across groups. The latter, unfortunately, often happens, for example, in disease-induced brain changes across clinical groups. Consequently, generative methods are usually inferior in prediction compared with the discriminative methods that target only the boundary of classes (such as Support Vector Machines (SVMs)). On the other hand, discriminative methods often encounter the difficulty of interpretation, which is critical in exploratory research aimed at both the understanding and the prediction. Thus, this paper is motivated by the advantages that can be gained by learning BNs that are both representative and discriminative. Different from the Bayesian network classifiers in [16], [17], [18] that address discrete variables, we learn discriminative BNs for continuous variables, which is often needed in many domains including neuroimaging-based brain research. Moreover, we learn for each class a BN with enhanced discrimination and maintain the BN repre-

sentation of each individual class for interpretation<sup>1</sup>. To achieve our goal, we propose two discriminative learning frameworks based on sparse Gaussian Bayesian network (SGBN).

In the first framework (termed KL-SGBN), we employ Fisher kernel [19] to link the generative models of SGBN to the discriminative classifiers of SVMs, and convert the SGBN parameter learning to Fisher kernel learning via maximizing a generalization bound of SVMs. The contributions of this framework include the following. i) By inducing Fisher kernel on SGBN models, we provide a way to obtain sample-specific SGBN-induced feature vectors that can be used by the discriminative classifiers such as SVMs. Through this, we bridge the generative models and the discriminative classifiers. ii) We propose a kernel learning approach to discriminatively learn the parameters of SGBNs by optimizing the performance of SVM. iii) As a by-product, the manipulation of Fisher kernel on SGBN provides a new way of variable selection for SGBNs. This framework has a computational advantage: through the mapping of Fisher kernel, the SGBN-induced feature vectors become linear functions of the SGBN parameters, which significantly simplifies the optimization problem in the learning process.

Unlike KL-SGBN where the discrimination is obtained by optimizing the classification performance of SVMs, in the second learning framework (termed MM-SGBN), we propose to optimize a criterion directly built upon the classification performance of SGBNs. The motivation is that optimizing the performance of SVMs may not necessarily guarantee an equivalent improvement on SGBNs when SGBNs are the goal of applications. The contribution of this framework is a max-margin based method to jointly learn SGBNs, one for each class, for both representation and discrimination.

In addition to the two discriminative SGBN learning frameworks, our contributions in this paper also include a new DAG constraint of SGBN based on topological ordering to ensure the validity of the graph. This new DAG constraint circumvents the awkward hard binarization of SGBN parameters in the process of optimization in [14], and simplifies the related optimization problems. This consequently makes it possible to optimize all the SGBN parameters together to avoid the influence of feature ordering encountered in the Block Coordinate Descent (BCD) optimization in [14]. Moreover, this new DAG constraint also circumvents the need for presetting candidate parent sets as in [17].

Although the discriminative learning frameworks proposed in this paper are general methods, we focus on their applications in neuroimaging analysis for the early

1. In this paper, we deal with the scenario that maintaining the BN representation of individual class is critical for the understanding of domain problems, such as the brain network models for the healthy and the diseased groups. However, it is not difficult to see our discriminative learning frameworks could be slightly modified to learn only a single BN as the existing “Bayesian network classifiers” for continuous variables. However, this deviates from our motivation and therefore is not unfolded in this paper.

diagnosis of mental diseases. A newly emerging field in the imaging-based neuroscience, called brain network analysis, attempts to model the brain as a complex network and study the interactions of brain regions via imaging-based features [20]. Such research is important because brain network change is often found to be a response of the brain to damages. Due to its causal semantics, BN has been employed to model the “effective connectivity” of the brain [14], [21], [22]. The directionality of the connections may disclose the pathways of how one brain region affects another. The discoveries may lend further credence to the evidence of causal relationship changes found in many mental diseases, such as the Alzheimer’s disease (AD) [23], [14], [24], [22], and uncover novel connectivity-based biomarkers for disease diagnosis. The proposed learning frameworks has been tested on multiple neuroimaging data sets. As demonstrated, our methods can significantly improve the discriminative power of the obtained SGBNs, as well as maintaining their representation capacity.

Early conference versions of this work were published in [25], [26]. In this paper, a significant extension has been made on the following aspects. First, we analyze the problems of the DAG constraint used in [25], [26], [14], and propose a new constraint with theoretically guaranteed DAG property to overcome those drawbacks. Second, we experimentally verify the new DAG constraint on benchmark Bayesian network data sets for network structure learning, and compare our method with another eight competing methods in the literature. Third, we update our two discriminative learning frameworks with the new DAG constraint and redo all the experiments in our early work [25], [26]. Fourth, we analyze the connections and differences between the two proposed discriminative learning frameworks, and conduct more comprehensive experiments to explore the characteristics of our frameworks with varied parameters, which has not been done in [25], [26].

The rest of the paper is organized as follows. Section 2 reviews SGBN and introduces the background of brain network analysis. Sections 3 elaborates two frameworks to learn discriminative and representative SGBNs from continuous data. Section 4 revisits the problem of the existing DAG constraint of SGBN, and proposes a new one based on topological ordering. The proposed two learning frameworks with the new DAG constraint are experimentally tested in Section 5. This paper is concluded in Section 6. The notations of symbols frequently occurring in this paper are summarized in Table 1.

## 2 BACKGROUND

To make this paper self-contained, we introduce the background for both the methodology and its application to brain network analysis. Please note that the methodology could be generalized to applications beyond the example given in this paper.

TABLE 1  
**Notation**

$x_i$	a random variable
$\mathbf{x}$	a sample of $m$ variables: $\mathbf{x} = [x_1, x_2, \dots, x_m]^\top$
$\mathbf{X}$	the data matrix of $n$ samples, $\mathbf{X} \in \mathbb{R}^{n \times m}$
$\mathbf{x}_{:,i}$	the $i$ -th row of $\mathbf{X}$ , representing a sample
$\mathbf{x}_{:,i}$	the $i$ -th column of $\mathbf{X}$ , representing the realization
$\Theta$	of the random variable $x_i$ on $n$ samples
$\Theta$	the parameters of a Gaussian Bayesian Network
$\Theta = [\theta_1, \dots, \theta_m]$ , $\Theta \in \mathbb{R}^{m \times m}$	
$\mathbf{Pa}_i$	a vector containing the parents of $x_i$
$\mathbf{PA}_i$	a matrix whose $j$ -th column represents a
	realization of $\mathbf{Pa}_i$ on the $j$ -th sample.
$\mathbf{G}$	an $m \times m$ matrix for BN: if there is a directed edge
	from $x_i$ to $x_j$ , $\mathbf{G}_{ij} = 1$ , otherwise $\mathbf{G}_{ij} = 0$
$\mathbf{P}$	an $m \times m$ matrix for BN: if there is a directed path
	from $x_i$ to $x_j$ , $\mathbf{P}_{ij} = 1$ , otherwise $\mathbf{P}_{ij} = 0$

### 2.1 Sparse Gaussian Bayesian Network (SGBN)

Because this paper is based on SGBN model, in the following, we review the fundamentals of SGBN in [14]. All the symbols are defined in Table 1.

A Bayesian network (BN)  $\mathcal{G}$  is a directed acyclic graph (DAG), i.e. there is no closed path within the graph. It expresses the factorization property of a joint distribution  $p(\mathbf{x}) = \prod_{i=1, \dots, m} p(x_i | \mathbf{Pa}_i)$ . The conditional probability  $p(x_i | \mathbf{Pa}_i)$  is assumed to follow a Gaussian distribution in Gaussian Bayesian Network (GBN). Each node  $x_i$  is regressed over its parent nodes  $\mathbf{Pa}_i$ :  $x_i = \boldsymbol{\theta}_i^\top \mathbf{Pa}_i + \varepsilon_i$ , where the vector  $\boldsymbol{\theta}_i$  is the regression coefficients, and  $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ . The structure of BN could be characterized by the  $m \times m$  matrix  $\mathbf{G}$  or  $\mathbf{P}$  (defined in Table 1), representing the edges / paths in the graph, respectively.

Identifying parent sets is critical for BN learning. Traditional methods often consist of two stages: the candidate parent sets are initially identified in the first stage and further pruned by some criteria in the second stage. A drawback arises that when a true parent is missing in the first stage, it will never be recovered in the second stage. The work in [14] proposed a different approach based on sparse GBN (SGBN), denoted as H-SGBN in this paper. In H-SGBN, each node  $x_i$  is regressed over all the other nodes, and its parent set is implicitly selected by the regression coefficients  $\boldsymbol{\theta}_i$  that are estimated through a constrained LASSO regression. The following optimization is solved in [14]:

$$\begin{aligned} \min_{\Theta} \sum_{i=1}^m & \| \mathbf{x}_{:,i} - \mathbf{PA}_i^\top \boldsymbol{\theta}_i \|_2^2 + \lambda_1 \| \boldsymbol{\theta}_i \|_1 \\ \text{s.t. } & \Theta_{ji} \times \mathbf{P}_{ij} = 0, \forall i, j = 1, \dots, m, \quad i \neq j. \end{aligned} \quad (2.1)$$

A challenge for BN learning is how to enforce the DAG property, i.e., avoiding directed cycles in the graph. A sufficient and necessary condition for being a DAG is proposed in [14], which requires  $\Theta_{ji} \times \mathbf{P}_{ij} = 0$  for all

$i$  and  $j$ . Note that  $\mathbf{P}_{ij}$  is an implicit function of  $\Theta_{ji}$  (i.e.,  $\mathbf{P} = \text{expm}(\Theta)$ , the matrix exponential function of  $\Theta$ , as in [14]). Eqn. (2.1) is difficult to solve. In [14], a block coordinate descent (BCD) method is employed to solve a LASSO-like problem efficiently. The whole  $\Theta$  is optimized column-wisely and iteratively. In each iteration  $t$ , only one column of  $\Theta$ , say  $\Theta_{:,j}$ , is optimized with  $\mathbf{P}$  fixed as  $\mathbf{P}^{(t-1)}$  in the last iteration. Then  $\Theta^{(t)}$ , with the updated column  $\Theta_{:,j}$ , is binarized to obtain  $\mathbf{G}^{(t)}$ , based on which,  $\mathbf{P}^{(t)}$  is recalculated by a Breadth-first search with  $x_i$  being the root node. The process is repeated until convergence. H-SGBN simultaneously obtains the structure and the parameters of an SGBN via learning  $\Theta$ , e.g., there is no edge  $i \rightarrow j$  if  $\Theta_{ij}$  is zero. It has been demonstrated to outperform the conventional two-stage methods in network edge recovery.

## 2.2 Brain Network Analysis

Neuroimaging modalities and analysis techniques can provide more sensitive and consistent measurements than traditional cognitive assessment for the early diagnosis of disease. Many mental disorders are found associated with subtle abnormalities distributed over the entire brain, rather than an individual brain region. The “distributive” nature of mental disorders suggests the alteration of interactions between brain regions (neuronal systems) and thus the necessity of studying the brain as a complex network. Brain networks are mathematically represented by graphical models, which can be constructed from neuroimaging data as follows. The brain images belonging to different subjects are first spatially aligned to a common stereotaxic space by affine or deformable transformation, and then partitioned into regions of interest (ROI), i.e., clusters of imaging voxels, using either data-driven methods or predefined brain atlas. A brain network is then modeled by a graph with each node corresponding to a brain region and each edge corresponding to the connectivity between regions. Brain network analysis studies three kinds of brain connectivity. In this paper, we focus on the “effective connectivity” that describes the influence one brain region exert upon another. Some early works in this field require a prior model of brain connectivity and most have only considered a small number ( $\leq 10$ ) of brain regions using techniques such as structural equation modeling [27] and dynamic causal modeling [28]. More recently, models such as BN and Granger Causality have also been introduced into this field. It is suggested that BN may have advantages over those lag-based methods for brain network analysis by an experimental fMRI study [21]. Among BN-related methods, it is worth noting that the work in [14] is completely data-driven, which recovers SGBN from more than 40 brain regions in fluorodeoxyglucose PET (FDG-PET) images. The method employs the strategy of sparsity constraint to handle relatively larger scale BN construction, and circumvents the traditional two-stage procedure for identifying parent

sets in many sparse BN learning methods [12], [10].

## 3 PROPOSED DISCRIMINATIVE LEARNING OF GENERATIVE SGBN

BN models are by definition generative models, focusing on how the data could be generated through an underlying process. In the context of neuroimage analysis, these models represent the effective brain connectivity of the given population. When used for classification, e.g., identifying AD patients from the healthy, the SGBN models are trained for each class separately. A new sample  $\mathbf{x}_i$  is then assigned to the class with the higher likelihood of SGBN. This may ignore some subtle but critical network differences that distinguish the classes. Therefore, we argue that the parameters of the generative model should be learned from the two classes jointly to keep the essential discrimination.

Integrating generative and discriminative models is an important research topic in machine learning. In [29], the related approaches are roughly divided into three categories: blending, staging and iterative methods. In blending methods, both the discriminative and the generative terms are incorporated into the same objective function. In staging methods, the discriminative model is trained on features provided by the generative model. In iterative methods, the generative and the discriminative models are trained iteratively to influence each other. In this paper, we propose two kinds of discriminative learning frameworks to achieve our goal. One is a staging method, called Fisher-kernel-induced discriminative learning (KL-SGBN). It extracts sample-based features from SGBN by Fisher kernel to optimize the classification performance of SVM. The other is a blending method, called max-margin-based discriminative learning (MM-SGBN). It directly optimizes the classification performance of SGBNs subject to maintaining SGBN’s representation capacity. The two frameworks are elaborated in the following sections, respectively.

### 3.1 Proposed Fisher-kernel-induced Discriminative Learning (KL-SGBN)

We first introduce the Fisher-kernel-induced discriminative learning of SGBN, i.e., KL-SGBN. The algorithm is illustrated in Fig. 1 and overviewed as follows. Given two classes in comparison, two SGBN models (with the parameters of  $\Theta_1$  and  $\Theta_2$ ) are learned, one for each individual class. The original samples are then mapped into the gradient space of the SGBN parameters  $\Theta_1$  and  $\Theta_2$  by Fisher kernel (Section 3.1.1). Through this mapping, each sample is represented by a new feature vector (called Fisher vector [19]) that is a function of  $\Theta = [\Theta_1, \Theta_2]$ . These sample-specific feature vectors are then fed into an SVM classifier to minimize its generalization errors by adjusting  $\Theta$  (Section 3.1.2). The obtained optimal  $\Theta_1^*$  and  $\Theta_2^*$  encode the discriminative information and therefore improve the original SGBNs.

In this way, we convert the discriminative learning of SGBN parameters to the discriminative learning of Fisher kernels.

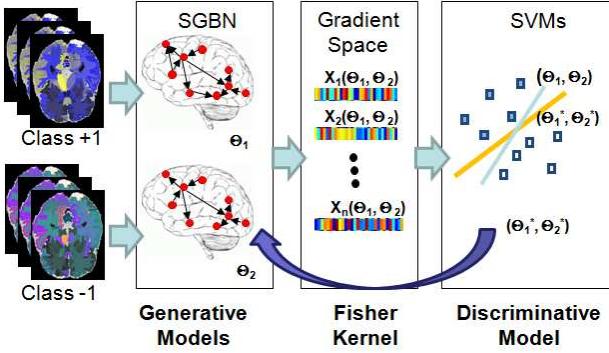


Fig. 1. Illustration of Fisher-kernel-induced Discriminative Learning.

### 3.1.1 Induction of Fisher vectors from SGBN

Below we introduce how to use Fisher kernel on SGBNs to obtain feature vectors required for kernel learning.

Fisher kernel [19] provides a way to compare samples induced by a generative model. It maps a sample to a feature vector in the gradient space of the model parameters. The intuition is that similar objects induce similar log-likelihood gradients of the model parameters. Fisher kernel is computed as  $K(\mathbf{x}, \mathbf{x}') = \mathbf{g}_x^\top \mathbf{U}^{-1} \mathbf{g}_{x'}$ , where the Fisher vector  $\mathbf{g}_x = \nabla_{\Theta} \log(p(\mathbf{x}|\Theta))$  describes the changing direction of parameters to better fit the model. The Fisher information metric  $\mathbf{U}$  weights the similarity measure, but is often set as an identity matrix in practice [19].

Fisher kernel has recently witnessed successful applications in image categorization [30], [31] for inducing feature vectors from Gaussian Mixture Model (GMM) of a visual vocabulary. Despite its success, in the applications above, Fisher kernel is mainly used as a feature extractor<sup>2</sup>. It has not been applied to learning the parameters of probability distributions before the early work of this paper in [25]. The advantage of learning discriminative Fisher kernel has also been confirmed by a recent study that maximizes the class separability [33] of samples based on Fisher kernel, which is developed with different context and different criteria from ours.

Following [14], we only consider  $\Theta$  as parameters and predefine  $\sigma$ . Let  $\mathcal{L}(\mathbf{x}|\Theta) = \log(p(\mathbf{x}|\Theta))$  denote the log-likelihood. Our Fisher vector for each sample  $\mathbf{x}$  is

$$\Phi_\Theta(\mathbf{x}) = [\nabla_{\Theta_1} \mathcal{L}(\mathbf{x}|\Theta_1)^\top, \nabla_{\Theta_2} \mathcal{L}(\mathbf{x}|\Theta_2)^\top]^\top,$$

where  $\Theta_1$  and  $\Theta_2$  are the parameters of the SGBNs for the two classes ( $y = 1, 2$ ), respectively. Recall that, using a BN, the probability  $p(\mathbf{x}|\Theta)$  can be factorized as

2. An exception [32] is discussed in “Generalization” in Section 3.3, which is published after our work [25].

$p(\mathbf{x}|\Theta) = \prod_{i=1, \dots, m} p(x_i|\mathbf{Pa}_i, \Theta_i)$ . Therefore, for GBN it can be shown that

$$\begin{aligned} \mathcal{L}(\mathbf{x}|\Theta) &= \sum_{i=1}^m \log p(x_i|\mathbf{Pa}_i, \Theta_i) \\ &= \sum_{i=1}^m \frac{-(x_i - \mathbf{Pa}_i^\top \Theta_i)^2}{2\sigma_i^2} - \log(2\pi\sqrt{\sigma_i}). \end{aligned} \quad (3.1)$$

Taking partial derivative over  $\Theta_i$ , we have

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathbf{x}|\Theta)}{\partial \Theta_i} &= -\frac{\mathbf{Pa}_i \mathbf{Pa}_i^\top}{\sigma_i^2} \Theta_i - \frac{x_i \mathbf{Pa}_i}{\sigma_i^2} \\ &\triangleq \mathbf{S}(x_i) \Theta_i + \mathbf{s}_0(x_i), \end{aligned} \quad (3.2)$$

where  $\mathbf{S}(x_i)$  is a squared matrix and  $\mathbf{s}_0(x_i)$  is a vector. As shown, both  $\mathbf{S}(x_i)$  and  $\mathbf{s}_0(x_i)$  are constant with respect to  $\Theta$ . Therefore, the Fisher vector  $\Phi_\Theta(\mathbf{x})$  is a linear function of  $\Theta$ . This simple form of  $\Phi_\Theta(\mathbf{x})$  significantly facilitates our further kernel learning.

### 3.1.2 Discriminative Fisher kernel learning via SVM

As each Fisher vector is a function of the SGBN parameters, discriminatively learning these parameters can thus be converted to learning discriminative Fisher kernels. We require that the learned SGBN models possess the following properties. Firstly, the Fisher vectors induced by the learned SGBN model should be well separated between classes. Secondly, the learned SGBN models should maintain reasonable capacity of representation. Thirdly, the learned SGBN models should not violate DAG.

We use the following strategies to achieve our goal. Firstly, to obtain a discriminative Fisher kernel, we jointly learn the parameters of SGBN and the separating hyper-plane of SVMs with Fisher kernel. Radius-margin bound, the upper bound of the Leave-One-Out error, is minimized to keep good generalization of the SVMs. Secondly, to maintain reasonable representation, we explicitly control the fitting (regression) errors of the learned model during optimization. Recall that GBN learns the network by minimizing the regression errors of each node over its parent nodes. Thirdly, we enforce the DAG constraint to ensure the validity of the graph. Our method is developed as follows.

In order to use radius-margin bound,  $\mathcal{L}_2$ -SVM with soft margin is employed [34]<sup>3</sup>, which optimizes

$$\begin{aligned} \min_{\mathbf{w}, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \xi^\top \xi \\ \text{s.t.} \quad & y_i (\mathbf{w}^\top \Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i. \end{aligned} \quad (3.3)$$

Following the convention in SVMs,  $\mathbf{w}$  is the normal of the separating plane,  $b$  the bias term,  $\xi$  the slack variables and  $C$  the regularization parameter. Here  $y_i$  is the class label of the  $i$ -th sample.  $\mathcal{L}_2$ -SVM can be rewritten as SVM with hard margin by slightly modifying the kernel

3. Radius-margin bound is rooted in hard-margin SVM.  $\mathcal{L}_2$ -SVM with soft-margin can be rewritten as SVM with hard margin.

$\mathbf{K} := \mathbf{K} + \mathbf{I}/C$ , where  $\mathbf{I}$  is identity matrix. For convenience, in the following, we redefine  $\mathbf{w} := [\mathbf{w}^\top \sqrt{C}\boldsymbol{\xi}^\top]^\top$  and  $\Phi(\mathbf{x}_i) := [\Phi^\top(\mathbf{x}_i) \ \mathbf{e}_i^\top y_i/\sqrt{C}]^\top$ . The vector  $\mathbf{e}_i$  has the value of 1 at the  $i$ -th element, and 0 elsewhere.

Incorporating radius information leads to solving

$$\begin{aligned} & \min_{\mathbf{w}} \frac{1}{2} R^2 \|\mathbf{w}\|_2^2 \\ \text{s.t. } & y_i(\mathbf{w}^\top \Phi(\mathbf{x}_i) + b) \geq 1, \quad \forall i, \end{aligned} \quad (3.4)$$

where  $R^2$  denotes the radius of Minimal Enclosing Ball (MEB). It has been observed that when the sample size is small, the estimation of  $R^2$  may become noisy and unstable [35]. Therefore, it has been proposed to use the trace of the total scatter matrix instead for such cases [35], [36]. We finally solve the following optimization problem:

$$\begin{aligned} & \min_{\theta, \mathbf{w}} \frac{1}{2} \text{tr}(\mathbf{S}_T) \|\mathbf{w}\|_2^2 \\ \text{s.t. } & y_i(\mathbf{w}^\top \Phi_\Theta(\mathbf{x}_i) + b) \geq 1, \quad \forall i \\ & h(\mathbf{X}_1, \Theta_1) \leq T_1, \quad h(\mathbf{X}_2, \Theta_2) \leq T_2, \\ & \Theta_1 \in \text{DAG}, \quad \Theta_2 \in \text{DAG}. \end{aligned} \quad (3.5)$$

Here  $\text{tr}(\mathbf{S}_T)$  is the trace of the total scatter matrix  $\mathbf{S}_T$ , where  $\mathbf{S}_T = \sum_{i=1}^n (\Phi(\mathbf{x}_i) - \mathbf{m})(\Phi(\mathbf{x}_i) - \mathbf{m})^\top$ , and  $\mathbf{m}$  is the mean of total  $n$  samples in the kernel-induced space. It can be shown that  $\text{tr}(\mathbf{S}_T) = \text{tr}(\mathbf{K}) - \mathbf{1}^\top \mathbf{K} \mathbf{1}/n$ , where  $\mathbf{1}$  denotes a vector whose elements are all 1, and  $\mathbf{K}$  the kernel matrix. Fisher vector  $\Phi_\Theta(\mathbf{x}_i)$  is obtained as in Section 3.1.1. The function  $h(\cdot)$  measures the squared fitting errors of the corresponding SGBNs for the data  $\mathbf{X}_1$  and  $\mathbf{X}_2$  from the two classes. It is defined as

$$h(\mathbf{X}, \Theta) = \sum_{i=1}^m \|\mathbf{x}_{:,i} - \mathbf{P} \mathbf{A}_i^\top \theta_i\|_2^2. \quad (3.6)$$

The two user-defined parameters  $T_1$  and  $T_2$  explicitly control the degree of fitting during the learning. Adding these constraints also avoids the scaling problem of  $\Theta$ .

The DAG constraint in H-SGBN could be employed to enforce the validity of the graph. However, here we adopt a new DAG constraint proposed in Section 4 due to its advantages over that of H-SGBN. The new DAG constraint employs a set of topological ordering variables  $(\mathbf{o}, \Upsilon)$  to guarantee DAG. It is a bilinear function of the ordering variables  $(\mathbf{o}, \Upsilon)$  and the SGBN parameters  $\Theta$ . An elaboration is given in Section 4. At the moment, let us temporarily skip the details of this DAG constraint and concentrate on the discriminative learning.

One possible approach for solving Eqn. (3.5) is to alternately optimize the separating hyperplane  $\mathbf{w}$  and the parameter  $\Theta$ . That is,

$$\begin{aligned} & \min_{\Theta, \mathbf{o}, \Upsilon} J(\Theta) \\ \text{s.t. } & h(\mathbf{X}_1, \Theta_1) \leq T_1, \quad h(\mathbf{X}_2, \Theta_2) \leq T_2, \\ & \Theta_1 \in \text{DAG}(\mathbf{o}_1, \Upsilon_1), \quad \Theta_2 \in \text{DAG}(\mathbf{o}_2, \Upsilon_2). \end{aligned} \quad (3.7)$$

---

**Algorithm 1** KL-SGBN: Discriminative Learning

---

**Input:** data  $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{n \times m}$ , label  $\mathbf{y} \in \mathbb{R}^{n \times 1}$

Denote  $\Theta = [\Theta_1, \Theta_2]$

Initialize  $\Theta^{(0)}, \mathbf{o}^{(0)}, \Upsilon^{(0)}$  by Algorithm 3 for each class.

Let  $\Theta^{(t-1)} = \Theta^{(0)}$ ,  $\mathbf{o}^{(t-1)} = \mathbf{o}^{(0)}$ ,  $\Upsilon^{(t-1)} = \Upsilon^{(0)}$

**repeat**

1. Compute  $\Phi_\Theta^{(t-1)}$  and  $\mathbf{K}_\Theta^{(t-1)}$  by Eqn. (3.2)
2. Compute  $\text{tr}(\mathbf{S}_T)^{(t-1)} = \text{tr}(\mathbf{K}_\Theta^{(t-1)}) - \mathbf{1}^\top \mathbf{K}_\Theta^{(t-1)} \mathbf{1}/n$
3. Solve  $J_0(\Theta^{(t-1)})$  and  $\alpha^*$  by Eqn. (3.9)
4.  $J(\Theta^{(t-1)}) = J_0(\Theta^{(t-1)}) \times \text{tr}(\mathbf{S}_T)^{(t-1)}$
6. Minimize Eqn. (3.7) with  $\alpha^*$  and obtain  $\Theta^{(t)}$ :
  - 6.1 Let  $\mathbf{o} = \mathbf{o}^{(t-1)}$ ,  $\Upsilon = \Upsilon^{(t-1)}$ , solve  $\Theta^{(t)}$  by Eqn. (3.7);
  - 6.2 Let  $\Theta = \Theta^{(t)}$ , solve  $\mathbf{o}^{(t)}$ ,  $\Upsilon^{(t)}$  by Eqn. (4.2).
7. Let  $\Theta^{(t-1)} = \Theta^{(t)}$ ,  $\mathbf{o}^{(t-1)} = \mathbf{o}^{(t)}$ ,  $\Upsilon^{(t-1)} = \Upsilon^{(t)}$

**until** convergence/max number of iterations

**Output:**  $\Theta^* = \Theta^{(t)}$

---

where

$$J(\Theta) = \min_{\mathbf{w}} \frac{1}{2} \text{tr}(\mathbf{S}_T) \|\mathbf{w}\|_2^2 \quad (3.8)$$

$$\text{s.t. } y_i(\mathbf{w}^\top \Phi_\Theta(\mathbf{x}_i) + b) \geq 1, \quad \forall i.$$

Note that for a given  $\Theta$ , the term  $\text{tr}(\mathbf{S}_T)$  is constant in Eqn. (3.8). Due to the strong duality in SVM optimization, we solve the term  $\|\mathbf{w}\|_2^2$  by

$$J_0(\Theta) = \max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j K_\Theta(\mathbf{x}_i, \mathbf{x}_j) \quad (3.9)$$

$$\text{s.t. } \sum_{i=1}^n \alpha_i y_i = 0, \quad \alpha_i \geq 0 \quad \forall i,$$

where  $\alpha_i$  is the Lagrangian multiplier and  $K_\Theta(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi_\Theta(\mathbf{x}_i), \Phi_\Theta(\mathbf{x}_j) \rangle$ .

As mentioned above, the DAG constraint is a bilinear function of  $(\mathbf{o}, \Upsilon)$  and  $\Theta$ . Many quadratic programming packages could be used to solve Eqn. (3.7). We use fmincon-SQP (sequential quadratic programming) in Matlab. Gradient information is required by many optimization algorithms (including fmincon-SQP) to speed up the line search. It is not difficult to find that the gradient of  $K_\Theta(\mathbf{x}_i, \mathbf{x}_j)$  is just a linear function of  $\Theta$ , making the evaluation of gradient  $\nabla_\Theta J$  easy. Our learning process is summarized in Algorithm 1.

### 3.2 Proposed Max-margin-based Discriminative Learning (MM-SGBN)

KL-SGBN introduces group discrimination into SGBNs by optimizing the performance of SVM classifiers with

SGBN-induced features. Although this leads to a relatively simple optimization problem, optimizing the performance of SVMs does not necessarily imply optimizing the discrimination of SGBNs. We believe that, the discrimination of SGBNs can be further improved if we *directly* optimize their (instead of SVMs') classification performance. Therefore we propose a new learning framework based on max-margin formulation directly built on SGBNs. We call this method MM-SGBN.

For binary classification, maximizing the minimum margin between two classes can be obtained by maximizing the minimum conditional likelihood ratio (MCLR) [18]:

$$\text{MCLR}(\Theta) = \min_{i=1}^n \frac{P(y_i|\mathbf{x}_i, \Theta_{y_i})}{P(\bar{y}_i|\mathbf{x}_i, \Theta_{\bar{y}_i})},$$

Without loss of generality,  $y_i$  and  $\bar{y}_i \in \{-1, 1\}$ , representing the true and false labels for the  $i$ -th sample, respectively. The parameter  $\Theta_{y_i} = \Theta_1$  if  $y_i = 1$ , or  $\Theta_{y_i} = \Theta_2$  if  $y_i = -1$ . We can see that MCLR identifies the most confusing sample whose probability of the true class assignment is close to or even less than that of the false class assignment. Hence, maximizing MCLR targets the maximal separation of the most confusing samples in the two classes. It is not difficult to see that MCLR can naturally handle multi-class case when replacing the denominator by the maximal probability induced by all false class assignments. Let  $\Theta = [\Theta_1, \Theta_2]$ . Taking log-likelihood of MCLR, we have

$$\begin{aligned} & \log \text{MCLR}(\Theta) \\ &= \min_{i=1}^n (\log p(\mathbf{x}_i|y_i, \Theta_{y_i}) - \log p(\mathbf{x}_i|\bar{y}_i, \Theta_{\bar{y}_i})) + \text{const}, \end{aligned} \quad (3.10)$$

where the prior probabilities of  $P(y_i)$  and  $P(\bar{y}_i)$  that are irrelevant to  $\Theta$  are absorbed into the constant term. Eqn. (3.10) can be shown to be a quadratic function of  $\Theta$  in the case of SGBN. In order to maximize MCLR, we require the difference of log-likelihood function in Eqn. (3.10) be larger than a margin for all samples,  $r$ , and maximize the margin  $r$ . To deal with hard separations, we employ a soft margin formulation as follows.

$$\min_{\Theta_1, \Theta_2, \xi_i, r, \mathbf{o}, \Upsilon} \lambda \sum_{i=1}^n \xi_i - r \quad (3.11)$$

$$\text{s.t. } y_i (\mathcal{L}(\Theta_1, \mathbf{x}_i) - \mathcal{L}(\Theta_2, \mathbf{x}_i)) \geq r - \xi_i, \quad \forall i \quad (3.11a)$$

$$\xi_i \geq 0, \quad r \geq 0, \quad (3.11b)$$

$$h(\mathbf{X}_1, \Theta_1) \leq T_1, \quad h(\mathbf{X}_2, \Theta_2) \leq T_2 \quad (3.11c)$$

$$\Theta_1 \in \text{DAG}(\mathbf{o}_1, \Upsilon_1), \quad \Theta_2 \in \text{DAG}(\mathbf{o}_2, \Upsilon_2) \quad (3.11d)$$

The constraints in (3.11a) enforce the likelihood of  $\mathbf{x}_i$  to its true class larger than that to its false class by a margin  $r$ . The variables  $\xi_i$  are slack variables indicating the intrusion of the margin. The function  $\mathcal{L}(\cdot)$  denotes the log-likelihood, defined in Eqn. (3.1). We require  $\mathcal{L}(\Theta_1, \mathbf{x}_i)$  larger than  $\mathcal{L}(\Theta_2, \mathbf{x}_i)$  when  $y_i = 1$ , and  $\mathcal{L}(\Theta_2, \mathbf{x}_i)$  larger than  $\mathcal{L}(\Theta_1, \mathbf{x}_i)$  when  $y_i = -1$ .

---

**Algorithm 2** MM-SGBN: Discriminative Learning

---

**Input:** data  $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{n \times m}$ , label  $\mathbf{y} \in \mathbb{R}^{n \times 1}$

Denote  $\Theta = [\Theta_1, \Theta_2]$

Initialize  $\Theta^{(0)}, \mathbf{o}^{(0)}, \Upsilon^{(0)}$  by Algorithm 3 for each class.

Fix  $\Theta = \Theta^{(0)}$  and estimate  $r^{(0)}$  and  $\xi_i^{(0)}$  by Eqn. (3.11)

only with the two constraints (3.11a) and (3.11b).

Initialize  $t = 1$ .

**repeat**

Step 1: Fixing  $\mathbf{o} = \mathbf{o}^{(t-1)}$  and  $\Upsilon = \Upsilon^{(t-1)}$ , optimize Eqn. (3.11) with the constraints (3.11a ~ 3.11c) to update  $\Theta^{(t)}$ ,  $r^{(t)}$  and  $\xi_i^{(t)}$ ;

Step 2: Fixing  $\Theta^{(t)}$ , optimize Eqn. (4.2) to update  $\mathbf{o}^{(t)}$  and  $\Upsilon^{(t)}$  to enforce DAG.

Let  $t = t + 1$

**until** convergence/max number of iterations

**Output:**  $\Theta^* = \Theta^{(t)}$

---

The constraints in (3.11c) control the fitting errors, same to that used in KL-SGBN, and the function  $h(\cdot)$  is defined in Eqn. (3.6).

The constraints in (3.11d) are the DAG constraint proposed in Section 4, Eqn. (4.1). To enforce the validity of DAG on both graphs, we introduce a set of order variables  $\mathbf{o} = \{o_1, o_2, \dots, o_m\}$  and  $\Upsilon$  for each class separately, and employ the constraints stated in Eqn. (4.1). Please refer to Eqn. (4.1) for details.

The optimization in Eqn. (3.11) can be solved iteratively by optimizing  $(\Theta, \xi_i, r)$  and  $(\mathbf{o}, \Upsilon)$  alternately, as summarized in Algorithm 2. In Step 1, we solve a linear objective function with  $n$  non-convex and two convex quadratic constraints by fmincon-SQP (sequential quadratic programming) in Matlab. In Step 2, we solve the linear programming by the package of CVX<sup>4</sup>.

It is worthy noting that, we learn an SGBN model for each individual class in order to meet the requirement of both interpretation and discrimination in exploratory research. For example, each SGBN may model the brain network of the healthy or the diseased class, as well as carrying the essential class discrimination. Both the network modelling and the discrimination are of interest in such cases. Our method is different from the conventional BN classifiers [16], [17], [18] that solely focus on classification. In those methods, only a single BN is learned to reflect the “difference” of the two classes. It does not model any individual class as our method does, and hence deviates from our purpose of both representing and discriminating brain networks. Moreover, the works in [16], [17], [18] cannot handle the continuous variables of brain imaging measures, and inherit the drawbacks of the traditional two-stage methods.

4. <http://cvxr.com/cvx/>

### 3.3 Discussion and Analysis

In the following, some issues regarding the two proposed discriminative learning frameworks are discussed.

**Classifiers.** The proposed discriminative learning frameworks produce a set of jointly learned SGBN models, one for each class. Based on these SGBN models, two kinds of classifiers can be constructed, i.e., the SGBN classifier and the SVM classifier. The SGBN classifier categorizes a sample by comparing the sample's likelihood according to each SGBN model. The SVM classifier is trained by the sample-specific Fisher vectors induced from the SGBN models. These two classifiers are tightly coupled by the underlying SGBN models. Specifically, more discriminative SGBN models directly lead to a better SGBN classifier, and can provide discriminative Fisher vectors to SVM for better classification. Rooted in this relationship, both the KL-SGBN and the MM-SGBN can improve the classification performance of these two classifiers simultaneously. Put simply, KL-SGBN explicitly optimizes the SVM classifier and in turn implicitly improves the SGBN classifier; while MM-SGBN explicitly optimizes the SGBN classifier, bringing an implicit improvement of the SVM classifier as well. When evaluating the discriminative power of the learned SGBN models by the SGBN classifier (a direct measurement), it is therefore expected that MM-SGBN can outperform KL-SGBN. However, KL-SGBN has some computational advantages and provides a new perspective to manipulate BN models, analyzed as follows.

**Computational Issues.** Compared with KL-SGBN, MM-SGBN requires to solve more complicated optimization problems, which may become problematic when the number of training samples increase. Let us compare Eqn. (3.7) for KL-SGBN and Eqn. (3.11) for MM-SGBN. For KL-SGBN, Eqn. (3.7) optimizes  $J(\Theta)$  with two convex quadratic constraints of data fitting and two DAG constraints, which are independent of the number of training samples  $n$ . The evaluation of  $J(\Theta)$  needs to solve an SVM-like problem in Eqn. (3.8), taking just  $n$  linear constraints of  $\Theta$ , which could be efficiently solved by off-the-shelf SVM packages. For MM-SGBN, in addition to the data fitting and DAG constraints as in Eqn. (3.7), the optimization problem in Eqn. (3.11) also has to satisfy  $n$  non-convex quadratic constraints. When  $n$  increases to a medium or large value, the optimization problem could be quite hard to solve.

**Edge Selection.** In addition to the discriminative learning of SGBN, the employment of Fisher kernel in KL-SGBN also provides a new perspective of edge selection for GBN. As introduced in Section 3.1.1, applying Fisher kernel on GBN produces sample-specific feature vectors whose component is the gradient of the log likelihood, i.e.,  $\frac{\partial \mathcal{L}(\mathbf{x}|\Theta)}{\partial \Theta_{ij}}$ . In other words, each feature now corresponds to an edge  $\Theta_{ij}$  in the SGBN. This makes it possible to convert the SGBN edge selection to a more traditional feature selection problem that has been well studied and has a large body of options in

the literature. Edge selection has been employed in our work to deal with the "small sample size" problem that is often encountered in medical applications. For example, it is common to have only 100 training samples but 3200 parameters (for SGBNs of 40 nodes from two classes) to learn in brain network analysis. To handle this issue, we keep using the whole  $\Theta$  for computing  $\mathbf{K}_\Theta$ , but only optimize a selected subset  $\Theta_s$ . There are many options to determine  $\Theta_s$ . We just compute the Fisher vector  $\Phi_\Theta$  for each sample, calculate the Pearson correlation between each component of  $\Phi_\Theta$  and the class labels on the training data, and select the top  $\theta_i$  with the highest correlations. To keep our problem simple, only the parameters associated with edges present in the graph are optimized to avoid the violation of DAG. It is remarkable that even this simple selection process has significantly improved the discrimination for both KL-SGBN and MM-SGBN. Note that this edge selection step is essentially different from that of the traditional two-stage methods. It is just an empirical method to handle the small sample size problem and will become *unnecessary* when sufficient training data are available. In contrast, identifying the candidate-parent sets is an indispensable step in two-stage methods to obtain computationally tractable solutions.

**Generalization.** We would like to point out that our learning framework of KL-SGBN could be easily generalized. It could be used to discriminatively learn the parameters of distributions other than that represented by GBN by just simply switching GBN to the target distribution, such as Gaussian Mixture Model (GMM). Indeed, this has been seen in [32], after our work [25]. However, as shown in this paper, the Fisher vector of GBN is a linear function of the model parameters, which significantly simplifies the learning problem. This favorable property may not be guaranteed with other distributions, including GMM.

## 4 PROPOSED DAG CONSTRAINT

In this section, we revisit H-SGBN and propose a new DAG constraint that could simplify the optimization problems in SGBN and its discriminative learning process as introduced in Sections 3.1 and 3.2.

### 4.1 H-SGBN Revisited

Recall that, the DAG constraint in H-SGBN (Section 2.1) utilizes the matrix  $\mathbf{P}$ , an implicit function of  $\Theta$ , which significantly complicates the optimization problem in Eqn. (2.1). In [14], for simplicity, in each optimization iteration,  $\mathbf{P}$  is first treated as a constant while optimizing  $\Theta$ , and then recalculated by searching on the binarized new  $\Theta$ . This hard binarization could introduce high discontinuity of  $\Theta$  into the optimization. Solving  $\Theta$  column-wisely by BCD may mitigate this problem since only one column of  $\Theta$  is changed in each iteration, inducing less discontinuity. However, we observe that the solution of BCD depends on which column of  $\Theta$

to be optimized first. In other words, if we randomly permute the ordering of features (the columns in  $\mathbf{X}$ ), we will obtain different SGBNs, which impairs the interpretability of the SGBN model. The optimization ordering matters because the matrix  $\mathbf{P}$  used in the DAG constraint changes with the ordering. This problem has been demonstrated in our experiment. Moreover, we find experimentally that if  $\mathbf{P}$  is solved as a whole instead of BCD, the optimization in Eqn. (2.1) will not converge but oscillate between some *non-DAG* solutions, possibly due to the high discontinuity mentioned above<sup>5</sup>. Early stop cannot help because no premature solution satisfies DAG. These optimization difficulties motivate our work of proposing a new DAG constraint that is much simpler for SGBN, as described below.

## 4.2 Proposed DAG constraint

It is known that, a BN is equivalent to a topological ordering (Page 362 in [37]). Therefore, we propose a new DAG constraint applicable to continuous variables with GBN based on this equivalence. With a few linear inequalities and variables separable from  $\Theta$ , the new DAG constraint significantly simplifies that used in [14]. Specifically, given a directed graph  $\mathcal{G}$  and the parameters  $\Theta$ , a real-valued order variable  $o_i$  is assigned to each node  $i$ , where  $0 \leq o_i \leq \Delta$ , and  $\Delta$  is a predefined arbitrary positive number. We propose a sufficient and necessary condition for  $\mathcal{G}$  to be DAG as in Proposition 1.

**Proposition 1.** Given a sparse Gaussian Bayesian Network parameterized by  $\Theta$  and its associated directed graph  $\mathcal{G}$  with  $m$  nodes, the graph  $\mathcal{G}$  is DAG if and only if there exist some  $o_i$  ( $i = 1, \dots, m$ ) and  $\Upsilon \in \mathbb{R}^{m \times m}$ , such that for arbitrary  $\Delta > 0$ , the following constraints are satisfied:

(4.1)

$$o_j - o_i \geq \frac{\Delta}{m} - \Upsilon_{ij}, \quad \forall i, j \in \{1, \dots, m\}, \quad i \neq j \quad (4.1a)$$

$$\Upsilon_{ij} \geq 0, \quad (4.1b)$$

$$\Upsilon_{ij} \times \Theta_{ij} = 0, \quad (4.1c)$$

$$\Delta \geq o_i \geq 0. \quad (4.1d)$$

Eqn.(4.1) leads to a topological ordering equivalent to DAG. The topological ordering means that if node  $j$  comes after node  $i$  in the ordering ( $o_j > o_i$ ), there cannot be a link from node  $j$  to node  $i$ , which guarantees the acyclicity. The proof of Proposition 1 is given in Appendix.

By Proposition 1, we remove the awkward hard binarization for computing  $\mathbf{P}$  in [14]. The inequalities of (4.1a, 4.1b, 4.1d) are linear to the ordering variables  $o_i$  and  $\Upsilon$ . The equation (4.1c) differs from the equation  $\Theta_{ji} \times P_{ij} = 0$  in [14] in that the variable  $\Upsilon_{ij}$  is now separable from  $\Theta_{ij}$  (while  $P_{ij}$  is not) and does not require the binarization of  $\Theta$ . This makes it tractable to

5. Please note that, solving  $\Theta$  column-wisely without updating  $\mathbf{P}$  in each iteration will only lead to non-DAG solutions

solve  $\Theta$  as a whole instead of BCD (to avoid the feature ordering problem).

It is worth noting that, provided  $\Theta$  is sparse, the number of constraints in Eqn. (4.1) could be significantly reduced. As can be seen, for any  $\Theta_{ij} = 0$ , as long as we set the corresponding  $\Upsilon_{ij}$  an arbitrary value greater than  $(\frac{1}{m} + 1)\Delta$ , all the conditions in Eqn. (4.1) will be automatically satisfied. Therefore, we only need to consider the constraints related to  $\Theta_{ij} \neq 0$ .

The idea of topological ordering is also used to design DAG constraint for the discrete variables in [38]. However, the work in [38] addresses the multinomial distribution of discrete variables, while here we target the Gaussian distribution of continuous variables. It is worthy noting that the constraint in [38] has to predefine candidate parent-node sets. Therefore, it inherits the drawbacks of the two-stage methods as pointed out in Section 1. This has been circumvented in our proposed DAG constraint for SGBN.

## 4.3 Estimation of SGBN from A Single Class

With our DAG constraint proposed in Eqn. (4.1), we could estimate SGBN from a single class as the initial solution to our discriminative learning of KL-SGBN or MM-SGBN. In particular, we optimize

$$\begin{aligned} \min_{\Theta, \mathbf{o}, \Upsilon} \sum_{i=1}^m & \| \mathbf{x}_{:,i} - \mathbf{P} \mathbf{A}_i^\top \boldsymbol{\theta}_i \|_2^2 + \lambda_1 \|\boldsymbol{\theta}_i\|_1 + \lambda_{dag} \boldsymbol{\epsilon}_i^\top |\boldsymbol{\theta}_i| \\ \text{s.t. } & o_j - o_i \geq \frac{\Delta}{m} - \Upsilon_{ij}, \quad \forall i, j \in \{1, \dots, m\}, \quad i \neq j \\ & 0 \leq o_i \leq \Delta, \quad \Upsilon_{ij} \geq 0, \end{aligned} \quad (4.2)$$

where  $\boldsymbol{\epsilon}_i$  is the  $i$ -th column of the matrix  $\Upsilon$ , and  $|\boldsymbol{\theta}_i|$  the component-wise absolute value of  $\boldsymbol{\theta}_i$ . This optimization problem is solved in an iterative way with two alternate steps in each iteration: i) optimize  $\mathbf{o}$  and  $\Upsilon$  (with  $\Theta$  fixed) and ii) optimize  $\Theta$  (with  $\mathbf{o}$  and  $\Upsilon$  fixed). This process is repeated until convergence. We call this proposed method OR-SGBN (Algorithm 3).

When the coefficient  $\lambda_{dag}$  is sufficiently large, the alternate optimization strategy of Eqn. (4.2) will converge to a DAG solution, as shown in Proposition 2 in Appendix. In practice, for numerical stability, we adopt a “warm start” strategy as in [14], that is, to gradually increase the values of  $\lambda_{dag}$  until the resulting  $\mathcal{G}$  becomes DAG. Specifically, we use a set of values of  $\lambda_{dag}$ :  $\lambda_{dag}^{(1)} < \lambda_{dag}^{(2)} < \dots < \lambda_{dag}^{(M)}$  to solve Eqn. (4.2) (Algorithm 3).

We use a bias variable  $x_0 = 1$  in the regression model to improve data fitting, thus  $x_i = [\boldsymbol{\theta}_i \ \boldsymbol{\theta}_0]^\top [\mathbf{P} \mathbf{A}_i \ 1] + \varepsilon_i$  ( $i > 1$ ). In the following part, we denote  $\boldsymbol{\theta}_i \triangleq [\boldsymbol{\theta}_i \ \boldsymbol{\theta}_0]$  and  $\mathbf{P} \mathbf{A}_i \triangleq [\mathbf{P} \mathbf{A}_i \ 1]$ . The bias term  $\boldsymbol{\theta}_0$  is learned together with other  $\boldsymbol{\theta}_i$ . This equals to introducing a bias node into the graph. It has no parent but is the parent of all the other nodes. If the original graph is a DAG, this does not cause the violation of DAG.

It is interesting yet challenging to analyze the network consistency of OR-SGBN. It is noted that Eqn. (4.2)

**Algorithm 3** OR-SGBN: SGBN from a single class

---

**Input:** data  $\mathbf{X} \in \mathbb{R}^{n \times m}$ 

 Initialize  $\Theta^{(0)}$  by least square fitting.

 Initialize  $\mathbf{o}^{(0)}$  and  $\Upsilon^{(0)}$  by solving Eqn. (4.2) with  $\Theta = \Theta^{(0)}$ .

 Let  $T = 1$ .

**repeat**

 Fixing  $\Upsilon = \Upsilon^{(T-1)}$  and  $\mathbf{o} = \mathbf{o}^{(T-1)}$ .

 Let  $t = 1$ ,  $\Theta^{(T-1,t=0)} = \Theta^{(T-1)}$ .

**for**  $\lambda_{dag} = \lambda_{dag}^{(1)}$  to  $\lambda_{dag}^{(M)}$  **do**

 Optimize Eqn. (4.2) with the initial solution  $\Theta^{(T-1,t-1)}$  to obtain  $\Theta^{(T-1,t)}$ .

 Let  $t=t+1$ .

**end for**

 Let  $\Theta^{(T)} = \Theta^{(T-1,M)}$ .

 Fixing  $\Theta^{(T)}$ , optimize Eqn. (4.2) to update  $\mathbf{o}^{(T)}$  and  $\Upsilon^{(T)}$  to enforce DAG.

 Let  $T = T + 1$ .

**until** convergence/max number of iterations

---

**Output:**  $\Theta^* = \Theta^{(T)}$ 

can be reorganized into a weighted LASSO problem, which can be conceptually linked to “adaptive LASSO” in the literature [39], [40], [41]. The analysis framework provided by these works is suggestive of promising strategies to analyze the network consistency for L1-penalized Gaussian networks. However, a complete treatment of this analysis for OR-SGBN requires a deep investigation. Considering the significant amount of the required workload and its importance, we will explore this problem in a separate paper in our future work.

## 5 EXPERIMENT

In this section, we investigate the properties of our proposed methods from three aspects: the DAG constraint, the discriminative learning process, and the resulting connectivity for brain network analysis. Four experiments are conducted, summarized in Table 2. The data sets and the experiments are elaborated as follows.

### 5.1 Neuroimaging Data Sets

We conduct our experiment on the publicly accessible ADNI [42] database to analyze brain effective connectivity for the Alzheimer’s disease. Three data sets are used from two imaging modalities of MRI and FDG-PET downloaded from ADNI. They are elaborated as follows.

**MRI** data set includes 120 T1-weighted MR images belonging to 50 mild cognitive impairment (MCI) patients and 70 normal controls (NC). These images are preprocessed by the typical procedure of intensity correction, skull stripping, and cerebellum removal. We segment the images into gray matter (GM), white matter (WM), and

cerebrospinal fluid (CSF) using the standard FSL<sup>6</sup> package, and parcellate them into 93 Region of Interest (ROI) based on an ROI atlas [43] after spatial normalization. The GM volume of each ROI is used as the imaging feature to characterize each network node. Forty ROIs are included in this study, following [14]. They have higher correlation with the disease and are mainly located in the temporal lobe and subcortical region. Studying brain morphology as a network can take the advantage of statistical tools from graph theory. Moreover, it has been reported that the covariation of gray matter morphology might be related to the anatomical connectivity [44]. PET data set includes 103 FDG-PET images (and their corresponding MR images) of 51 AD patients and 52 NC. The MR images belonging to different subjects are co-registered and partitioned into ROIs as before. The ROI partitions are copied onto their corresponding PET images by a rigid transformation. The average tracer uptakes within each ROI is used as the imaging feature to characterize each network node. Forty ROIs discriminative to the disease are used in the study. The retention of tracer in FDG-PET is analogous to the glucose uptake, thus reflecting the tissue metabolic activity.

**MRI-II** data set is similar to the MRI data set but using 40 different ROIs covering the typical brain regions spread over the frontal, parietal, occipital and temporal lobes.

We randomly partition each data set into 30 groups of training-test pairs. Each group includes 80 training and 40 test samples in MRI and MRI-II, or 60 training and 43 test samples in PET.

### 5.2 DAG Constraint

With our proposed DAG constraint, the SGBN model for an individual class can be learned with all the parameters  $\Theta$  optimized together (OR-SGBN), instead of column-wisely as did in [14], [25], [26]. To explore the properties of our DAG constraint, we test three experimental configurations, namely, OR-SGBN (WHOLE), H-SGBN (BCD) and H-SGBN (WHOLE). The word in the parenthesis is used to explicitly indicate whether the parameters  $\Theta$  are optimized together (WHOLE) or column-wisely (BCD). OR-SGBN (WHOLE) is our SGBN learning method for a single class in Algorithm 3, implemented with the package of CVX. H-SGBN (BCD) is the column-wise method in [14] and implemented with the code downloaded from the authors’ website. H-SGBN (WHOLE) is our attempt to optimize  $\Theta$  together for the objective function of H-SGBN in [14], which is implemented with the package of CVX<sup>7</sup>. The same  $\Theta$  that is computed by a sparse least square fitting of the training set is provided to all the methods to initialize the optimizations. The “warm-start” strategy is applied wherever applicable in all methods.

6. <http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/>

7. The optimization problem is solved by a series of convex subproblems.

**TABLE 2**  
***Summary of Experiment Purpose***

Experiment	Test Subject	Purpose
Exp-I (Sec. 5.2)	DAG constraint	Test the invariance of solution to feature ordering
Exp-II (Sec. 5.2)	DAG constraint	Test the ability of network structure recovery
Exp-III (Sec. 5.3)	discriminative learning	Test the improvement of discriminative power of SGBN models
Exp-IV (Sec. 5.4)	brain network analysis	Investigate the learned brain connectivity patterns

It is found that when solving all  $\Theta$  as a whole, H-SGBN (WHOLE) that uses the DAG constraint in [14] does not converge: the optimization is trapped to oscillate between a few solutions that are not DAG. Therefore, from now on, we only consider H-SGBN (BCD) and OR-SGBN (WHOLE).

**Exp-I.** In this experiment, we compare the solutions of OR-SGBN (WHOLE) and H-SGBN (BCD) with respect to the change of feature ordering. To do that, for the neuroimaging data sets, we randomly permute the feature ordering for 100 times. The estimated  $\Theta$  of the resulting 100 SGBNs are re-arranged according to the initial feature ordering and then averaged as in Fig. 2. As shown, the averaged result from OR-SGBN (WHOLE) (Fig. 2 (d)) is almost identical to the result using the original feature ordering (Fig. 2 (c)), reflecting its robustness to feature ordering. In contrast, H-SGBN (BCD) generates SGBNs with large variations when the feature ordering changes ((Fig. 2 (a) versus (b)). To give a quantitative evaluation, the Euclidean distance and the correlation between the averaged  $\Theta$  and the original  $\Theta$  are presented in Table 3. Consistently, the solutions from OR-SGBN (WHOLE) are much less affected by the ordering permutation, indicating the advantage of solving  $\Theta$  as a whole via the proposed DAG constraint.

**TABLE 3**  
***Quantitative Analysis of  $\Theta$  for the random permutation of feature ordering (between the original and the averaged  $\Theta$ )***

		Distance	Correlation (R)
OR-SGBN (WHOLE)	$\Theta_1$	0.08	0.9996
	$\Theta_2$	0.18	0.9981
H-SGBN (BCD)	$\Theta_1$	1.91	0.6828
	$\Theta_2$	2.06	0.6396

**Exp-II.** In this experiment, we test the ability of OR-SGBN (WHOLE) at identifying network structures from data. Since no ground-truth is available for the three neuroimaging data sets due to the unknown mechanism of the disease, we conduct experiments on nine benchmark network data sets mostly coming from the Bayesian Network Repository [45] as was done in the literature [12], [46]. The nine benchmark data sets are: Factors (27 nodes, 68 arcs), Alarm (37 nodes, 46 arcs),

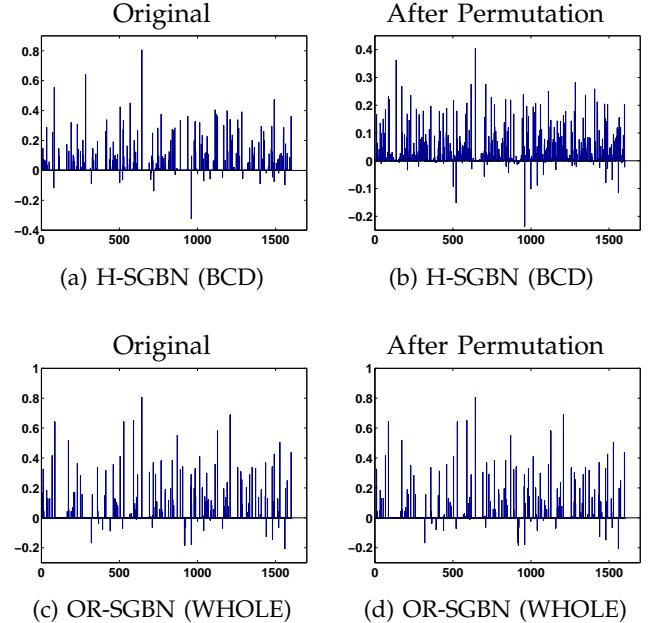


Fig. 2. One example of the estimated parameter  $\Theta$  for the MCI class (reshaped as a long vector) with regard to the random permutation of the feature ordering. Quantitative measurements of the changes are given in Table 3.

Barley (48 nodes, 84 arcs), Carpo (61 nodes, 74 arcs), Chain (7 nodes, 6 arcs), Hailfinder (56 nodes, 66 arcs), Insurance (27 nodes, 52 arcs), Mildew (35 nodes, 46 arcs) and Water (32 nodes, 66 arcs). We compare the OR-SGBN (WHOLE) with another eight BN learning methods, including L1MB [12], GS [47], TC and its variant TC-bw [13] and three variants of IAMB [48]. The experiment is repeated for 50 simulations. In each simulation, for each network, we randomly sample 1000 samples from  $\pm \text{Uniform}(0.5, 1)$  for the regression coefficients of each variable on its parents. The parameters of the eight methods to be compared are set according to [14]. A predefined  $\lambda$  that controls the sparsity of OR-SGBN is uniformly applied to all the nine data sets, which simply brings the number of the resulting edges to a reasonable range <sup>8</sup>. We use the first stage estimate of L1MB as the initial solution of OR-SGBN. Table 4 shows the total numbers of mis-identified edges (including both the false and the missing edges), while Table 5 shows the numbers of falsely identified edges (false positive).

<sup>8</sup>. The Bayesian Information Criterion is used to select  $\lambda$  in [14]. However, it did not behave well in our experiment.

In addition, Table 6 lists the numbers of falsely identified PDAG structures. PDAG structures are statistically indistinguishable structures, i.e., representing the same statistical dependency. The PDAG of BN is obtained by the method in [49]. From Tables 4 ~ 6, it can be seen that OR-SGBN shows significantly smaller errors on six data sets (Factors, Alarm, Barley, Carpo, Hailfinder and Insurance) in identifying both edges and PDAG structures. For the data sets of Mildew and Water, OR-SGBN performs similarly to the other methods. It only performs relatively inferior on Chain. This experiment demonstrates that the proposed DAG constraint for SGBN can perform effectively for BN structure identification. Its relatively low risk of mis-edge identification is a favorable property for exploratory research.

### 5.3 Comparison of Discrimination

After testing the effectiveness of the proposed DAG constraint, we now investigate the theme of this paper: the discriminative learning frameworks. We consider two kinds of classifiers: i) the SGBN classifier (with two SGBN models, one for each class), and ii) the SVM classifier learned by the Fisher vectors induced from the SGBN models.

**Exp-III.** In this experiment, we test whether our learning methods (KL-SGBN and MM-SGBN) can improve the discriminative power on both kinds of classifiers for the real neuroimaging data sets. The initial SGBN models are obtained by our proposed OR-SGBN (WHOLE), since it has been shown more robust to feature ordering than H-SGBN as above. For the SGBN classifier, assuming equal prior, we assign a test sample to the class with a higher likelihood. For the SVM classifier, we use  $\mathcal{L}_2$ -SVM with Fisher kernels. In order to maintain representation capability, we allow maximal 1% additional squared fitting errors (that is,  $T_i = 1.01 \times T_{i0}$ , ( $i = 1, 2$ ), where  $T_{i0}$  is the squared fitting error of the initial solution) to be introduced during the learning process of KL-SGBN or MM-SGBN.

The test accuracies are averaged over the 30 randomly partitioned training-test groups. The classification performances of SGBN and SVM classifiers are evaluated with the varied parameter  $\lambda$  that controls the sparsity level and the number of edges optimized in the learning process in Fig. 3. The results of our proposed KL-SGBN and MM-SGBN are plotted by the green and the red lines, respectively. The results of the individually learned OR-SGBN and H-SGBN are plotted by the blue and the black lines, respectively. The top two rows in Fig. 3 correspond to the results from the SGBN classifiers, while the bottom two rows correspond to those from the SVM classifiers. From Fig. 3, we have the following observations.

i) Both KL-SGBN and MM-SGBN can significantly improve the discriminative power of the individually learned SGBNs (Fig. 3, the top two rows), as well as their associated SVM classifiers (Fig. 3, the third row).

Such improvements are consistent over the three neuroimaging data sets and different parameter settings, and could reach the significant increases of 10% ~ 20% on most occasions. When the network becomes more sparse, the classification performance of the individually learned SGBNs (H-SGBN and OR-SGBN) drops significantly possibly due to the insufficient modeling of data. However, under such circumstances, KL-SGBN and MM-SGBN can still maintain high classification accuracies, which may indicate the necessity and effectiveness of the discriminative learning in classification scenarios.

ii) When using SGBN classifiers, for all the three data sets, MM-SGBN consistently achieves higher test accuracies at all sparsity levels (Fig. 3, the first row) with different numbers of optimized edges than KL-SGBN (Fig. 3, the second row). The advantage of MM-SGBN over KL-SGBN comes from explicitly optimizing the discriminative power of SGBNs, instead of getting help from optimizing the performance of SVM on SGBN-induced features.

iii) When using SVM classifiers, the SVM built upon KL-SGBN-induced features performs slightly better than that built upon MM-SGBN-induced features at all sparsity levels (Fig. 3, the third row). This is expected since KL-SGBN optimizes the performance of its associated SVM classifier.

iv) When cross-referencing the first and the third rows in Fig. 3, it is noticed that SVM classifiers in general perform worse than the discriminatively learned SGBN classifiers. These may be because our Fisher vectors have very high dimensionality, which causes serious overfitting of data in SVM classifiers. Such situation might be somewhat improved for SGBN-classifiers since the simple Gaussian model may “regularize” the model fitting. Based on this assumption, we further select a number of leading features from Fisher vectors by computing the Pearson correlation of the features and the labels, and use the selected features to construct the Fisher kernel for the SVM classifiers. As shown in the fourth row of Fig. 3, the simple feature selection step can further significantly improve the classification performance of the Fisher-kernel based SVM.

v) The individually learned OR-SGBN and H-SGBN perform similarly for classification. However, as mentioned above, OR-SGBN has an additional advantage of being invariant to the feature ordering.

vi) Recall that these improvement on discrimination are achieved with no more than 1% increase of squared fitting errors, which is explicitly controlled through the user defined parameters  $T_1$  and  $T_2$ . Note that the rate of 1% is application dependent. More tolerance of fitting errors can potentially bring more discrimination.

### 5.4 Comparison of Connectivity

We also conduct an investigation to gain some insights into the learned brain networks for the diseased and the healthy populations, respectively.

TABLE 4  
**Total errors (number of both false and missing edges, averaged on 50 simulations) on benchmark networks**

	L1MB	GS	TC-bw	TC	IAMB	IAMB1	IAMB2	IAMB3	OR-SGBN
Factors	101.48	104.50	102.90	103.02	103.14	103.30	103.14	103.14	<b>54.82</b>
Alarm	56.58	59.30	57.76	60.60	61.76	59.16	61.76	61.76	<b>44.40</b>
Barley	113.24	114.70	114.38	122.78	123.80	109.92	123.80	123.80	<b>99.26</b>
Carpo	125.74	131.72	131.18	133.16	132.76	132.90	132.76	132.76	<b>25.58</b>
Chain	5.32	4.88	5.50	<b>4.42</b>	4.70	5.00	4.70	4.70	7.04
Hailfinder	91.50	94.94	96.18	99.02	103.10	92.74	103.10	103.10	<b>57.04</b>
Insurance	74.78	74.64	73.74	76.30	78.78	73.04	78.78	78.78	<b>59.04</b>
Mildew	60.86	60.74	59.66	63.80	68.46	92.74	103.10	103.10	<b>57.04</b>
Water	92.86	94.04	90.24	97.16	102.70	<b>90.06</b>	102.70	102.70	93.08

TABLE 5  
**Number of falsely identified edges (averaged on 50 simulations) on benchmark networks**

	L1MB	GS	TC-bw	TC	IAMB	IAMB1	IAMB2	IAMB3	OR-SGBN
Factors	47.66	50.74	49.40	49.74	50.28	49.70	50.28	50.28	<b>17.70</b>
Alarm	36.04	37.72	36.86	39.24	40.96	37.30	40.96	40.96	<b>23.14</b>
Barley	71.70	72.30	72.60	80.76	82.96	69.76	82.96	82.96	<b>48.70</b>
Carpo	71.96	76.30	75.14	77.38	77.18	77.36	77.18	77.18	<b>14.56</b>
Chain	2.66	2.44	2.76	<b>2.22</b>	2.36	2.50	2.36	2.36	3.52
Hailfinder	60.00	62.04	63.16	65.42	66.40	60.90	66.40	66.40	<b>28.66</b>
Insurance	42.80	42.16	41.72	44.06	48.08	41.42	48.08	48.08	<b>31.20</b>
Mildew	46.22	46.46	45.46	49.82	52.48	44.82	52.48	52.48	<b>33.86</b>
Water	64.52	65.02	63.70	68.06	74.22	63.48	74.22	74.22	<b>46.74</b>

TABLE 6  
**Number of falsely identified P-DAG structures (averaged on 50 simulations) on benchmark networks**

	L1MB	GS	TC-bw	TC	IAMB	IAMB1	IAMB2	IAMB3	OR-SGBN
Factors	107.20	109.54	108.96	108.84	109.22	108.84	109.22	109.22	<b>63.40</b>
Alarm	61.74	64.08	62.54	65.34	66.82	63.98	66.82	66.82	<b>51.02</b>
Barley	120.54	122.26	121.38	130.04	131.24	116.92	131.24	131.24	<b>105.50</b>
Carpo	129.02	135.34	134.78	136.92	136.74	136.22	136.74	136.74	<b>33.74</b>
Chain	5.96	5.54	6.12	<b>5.16</b>	5.30	5.66	5.30	5.30	7.42
Hailfinder	103.72	106.08	107.56	110.04	113.44	104.86	113.44	113.44	<b>63.76</b>
Insurance	81.58	81.68	81.44	83.70	86.60	80.66	86.60	86.60	68.26
Mildew	61.68	61.32	60.34	64.48	69.30	<b>58.08</b>	69.30	69.30	67.24
Water	96.34	97.46	93.80	100.38	106.14	<b>93.60</b>	106.14	106.14	94.52

**Exp-IV.** In this experiment, we visualize the learned brain networks and compare the connectivity patterns obtained by different methods and from different populations. MRI-II data set is used for this study since it covers regions spread over the four lobes of brain.

The structures of the brain networks recovered from NC and MCI groups are displayed in Fig. 4 by using H-SGBN (BCD) and OR-SGBN (WHOLE), respectively. The network structure is obtained by thresholding the edge weights  $\Theta$  with a cutoff value of 0.01 [14]. Each row  $i$  represents the effective connections (dark dots) starting from the node  $i$ , and each column  $j$  represents the effective connections ending at the node  $j$ . Note that, due to the different optimization problems involved, the same parameter  $\lambda$  leads to different sparsity levels for H-SGBN and OR-SGBN. However, for a given method, different  $\lambda$  values do not change the major structures of the resulting networks.

In Fig. 4, it is noticed that H-SGBN (BCD) usually generates more connections in the upper triangle of the graphs even when we randomly permute the nodes. We

suspect that this is caused by the column-wise optimization. The parameters  $\theta_i$  (corresponding to the columns in the graph) optimized at the early stage tend to be made more sparse than those optimized later in order to satisfy the DAG constraint. This phenomenon is not observed in OR-SGBN (WHOLE) that is used to initialize the discriminative learning.

Let us focus on OR-SGBN. Compared with H-SGBN, OR-SGBN has an additional bias node corresponding to the last row and column in Fig. 4. Visualizing  $\Theta$  can provide rich information for medical analysis. Here we just list a few observations as examples. With the same  $\lambda$ , OR-SGBN produces 183 edges for NC, and 145 edges for MCI. Such loss of connectivity also happens at the temporal lobe (24% loss) for MCI. The temporal lobe (and some subcortical structures) is known to play a very important role in the progression of AD. The loss of connectivity in this region has been well-documented in wide AD-related studies [50], [51], [14]. In Fig. 4, we also observe an increase of connectivity (the left bottom corner in the figure) between the frontal and the temporal

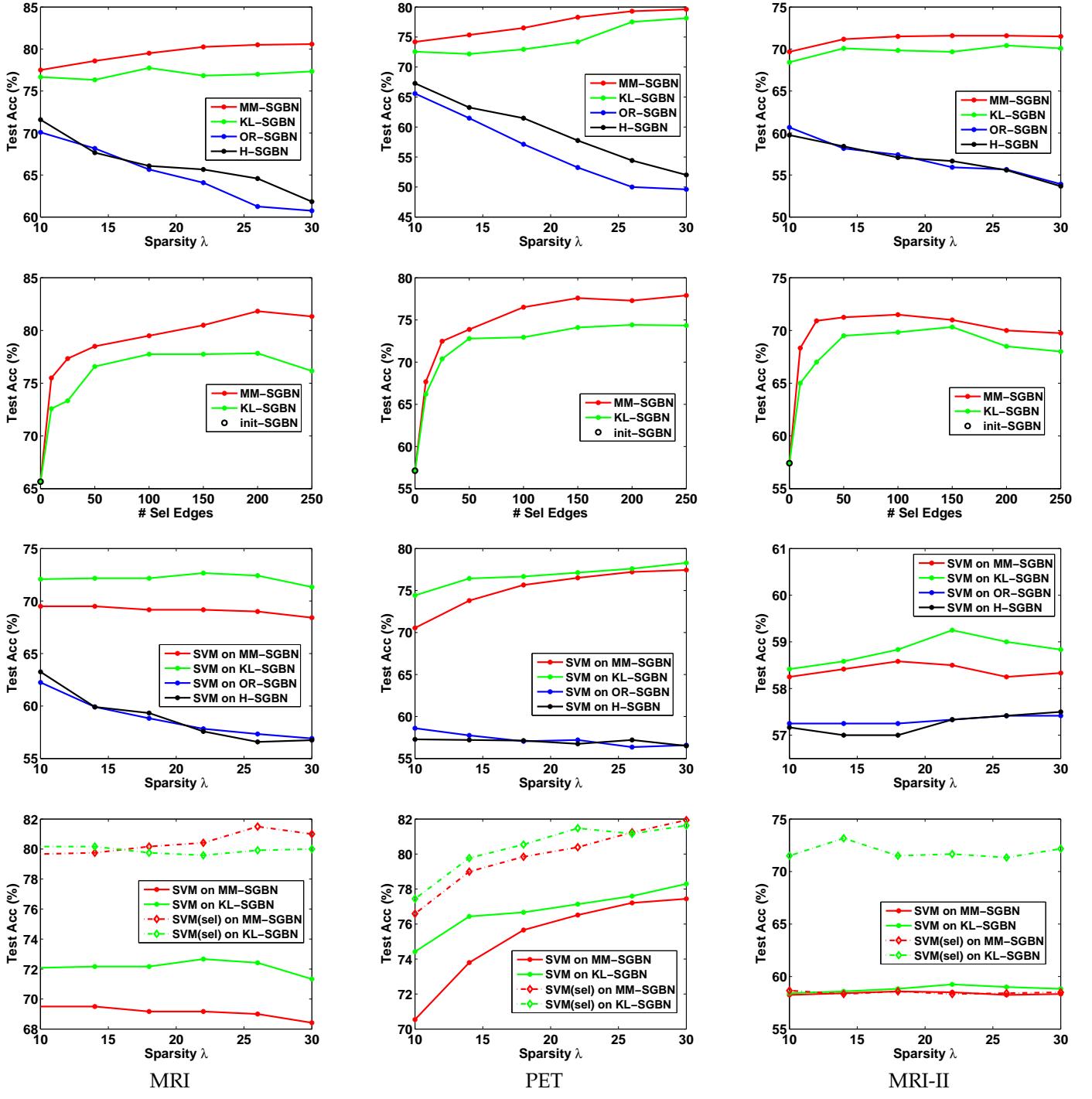


Fig. 3. Comparison of classification accuracies on data sets of MRI (the left column), PET (the middle column) and MRI-II (the right column). The top two rows correspond to the test accuracies obtained by the learned SGBNs. The first row shows the test accuracies varied with the sparsity levels (i.e., the parameter  $\lambda$ ). The second row shows the test accuracies varied with the number of edges (denoted as “#Sel Edges” in the figure) optimized in discriminative learning. The bottom two rows correspond to the test accuracies obtained by SVMs using the SGBN-induced Fisher vectors either in full length (the third row) or with (100) selected components (the fourth row).

lobe in MCI. Some study [52] mentioned that the frontal lobe may have connectivity increase at the early stage of AD as a compensation of cognitive functions for the patients. Moreover, significant directionality changes are also found for the left (node 35) and the right (node 38)

hippocampi, an important structure among the earliest ones affected by AD. Both hippocampi have reduced incoming connectivity (communications dominated by other nodes) but increased outgoing connectivity (communications dominated by themselves) in MCI. Please

note that the above observations could be influenced by the factors such as the limited number of data, the degree of disease progression and the imaging modality used in this study. More reliable medical analysis should be validated on larger data sets and worth further exploration, which is, however, beyond the scope of this paper.

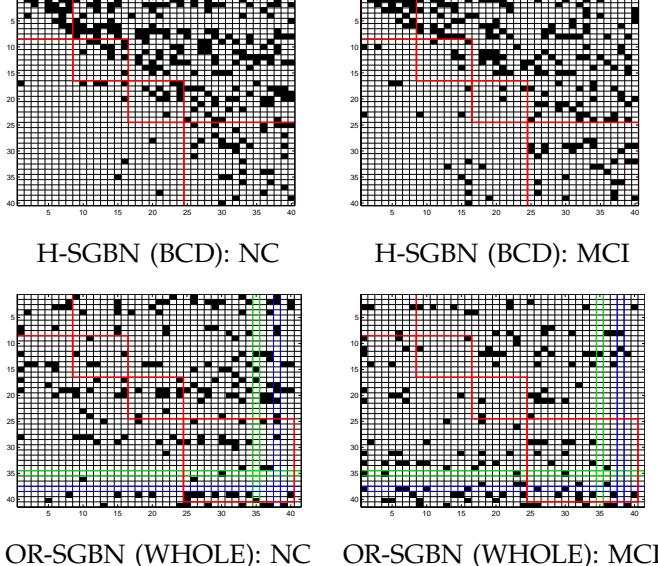


Fig. 4. *Visualization of connectivities for MRI-II. The four red boxes correspond to the frontal, parietal, occipital and temporal (including subcortical regions) lobes of the brain. The green row (Row 35) and column (Col 35) correspond to the left hippocampus while the blue ones (Row 38 and Col 38) correspond to the right hippocampus.*

To illustrate the difference of edge weights learned by KL-SGBN and MM-SGBN, an example of 30 edge weight changes (from the initial OR-SGBN) learned by these two methods is given in Fig. 5, where the SGBN networks from the two classes are vectorized and concatenated as  $x$ -axis. As shown, the signs of weight changes are quite similar in both methods. The most significant difference is that, MM-SGBN gives negative weight changes to the bias node of the left Amygdala and the right Parahippocampus (red lines in Fig. 5) while KL-SGBN gives them positive weight changes. The adjustment of edge weights leads to 10% increase of test accuracy for MM-SGBN in this example.

## 6 CONCLUSION

In this paper, we focus on the discriminative learning of Bayesian network for continuous variables, especially for neuroimaging data. Two discriminative learning frameworks are proposed to achieve this goal, i.e., KL-SGBN improves the performance of SVM classifiers based on SGBN-induced features, and MM-SGBN explicitly optimizes an SGBN-based criterion for classification. We demonstrate how to utilize Fisher-kernel to bridge the generative methods of SGBN and the discriminative classifiers of SVM, and how to embed the max-margin

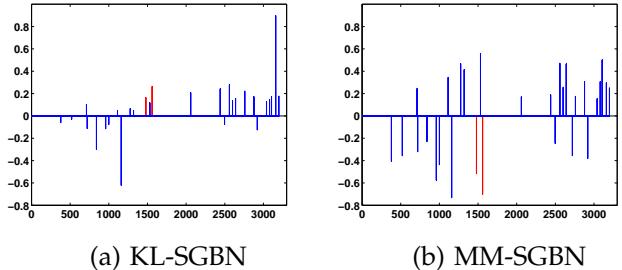


Fig. 5. *An example: change of edge weights learned by KL-SGBN and MM-SGBN*

criterion into SGBN for discriminative learning. The optimization problems are analyzed in details, and the advantages and disadvantages of the proposed methods are discussed. Moreover, a new DAG constraint is proposed to ensure the validity of the graph with theoretical guarantee and validation on the benchmark data. We apply the proposed methods to modeling brain effective connectivity for early AD prediction. Significant improvements have been observed in the discriminative power of both the SGBN models and the associated SVM classifiers, without sacrificing much representation power.

## REFERENCES

- [1] T. Verma and J. Pearl, "Equivalence and synthesis of causal models," *Uncertainty in Artificial Intelligence*, vol. 6, pp. 255–268, 1991.
- [2] P. Spirtes and C. Glymour, "An algorithm for fast recovery of sparse causal graphs," *Social Science Computer Review*, vol. 9, no. 1, pp. 62–72, 1991.
- [3] A. Fast, *Learning the structure of Bayesian networks with constraint satisfaction*. Ph.D thesis, University of Massachusetts Amherst, 2010.
- [4] M. Scutari, "Bayesian network constraint-based structure learning algorithms: Parallel and optimised implementations in the bnlearn r package," *CoRR*, vol. abs/1406.7648, 2014.
- [5] N. Friedman and D. Koller, "Being bayesian about network structure - bayesian approach to structure discovery in bayesian networks," *Machine Learning*, vol. 50, no. 1-2, pp. 95–125, 2003.
- [6] M. Koivisto and K. Sood, "Exact bayesian structure discovery in bayesian networks," *Journal of Machine Learning Research*, vol. 5, pp. 549–573, 2004.
- [7] D. Geiger and D. Heckerman, "Learning gaussian networks," *CoRR*, vol. abs/1302.6808, 2013.
- [8] J. Suzuki, "A construction of bayesian networks from databases based on an mdl principle," in *UAI*, 1993, pp. 266–273.
- [9] S. Acid and L. Campos, "Searching for bayesian network structures in the space of restricted acyclic partially directed graphs," *Journal of Artificial Intelligence Research*, vol. 18, pp. 445–490, 2003.
- [10] I. Tsamardinos, L. Brown, and C. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine Learning*, vol. 65, no. 1, pp. 31–78, 2006.
- [11] J. L. M. Jose A Gamez and J. M. Puerta., "Learning bayesian networks by hill climbing: Efficient methods based on progressive restriction of the neighborhood," *Data Mining and Knowledge Discovery*, vol. 22, no. 1-2, pp. 106–148, 2011.
- [12] M. Schmidt, A. Niculescu-Mizil, and K. Murphy, "Learning graphical model structures using l1-regularization paths," in *AAAI*, 2007.
- [13] J. Pellet and A. Elisseeff, "Using markov blankets for causal structure learning," *JMLR*, vol. 9, pp. 1295–1342, 2008.

- [14] S. Huang, J. Li, J. Ye, A. Fleisher, K. Chen, T. Wu, and E. Reiman, "A sparse structure learning algorithm for gaussian bayesian network identification from high-dimensional data," *IEEE TPAMI*, vol. 35, no. 6, pp. 1328–1342, 2013.
- [15] J. Xiang and S. Kim, "A\* lasso for learning a sparse bayesian network structure for continuous variables," in *NIPS*, 2013, pp. 2418–2426.
- [16] F. Pernkopf and J. Bilmes, "Efficient heuristics for discriminative structure learning of bayesian network classifiers," *JMLR*, vol. 11, pp. 2323–2360, 2010.
- [17] F. Pernkopf, M. Wohlmayr, and S. Tschiatschek, "Maximum margin bayesian network classifiers," *IEEE TPAMI*, vol. 34, no. 3, pp. 521–532, 2012.
- [18] Y. Guo, D. Wilkinson, and D. Schuurmans, "Maximum margin bayesian networks," in *UAI*, 2005.
- [19] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *NIPS*, 1998.
- [20] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nat Rev Neurosci*, vol. 10, no. 3, pp. 186–198, 2009.
- [21] S. Smith, K. Miller, G. Khorshidi, M. Webster, C. Beckmann, T. Nichols, J. Ramsey, and M. Woolrich, "Network modelling methods for fmri," *Neuroimage*, vol. 54, no. 2, pp. 875–891, 2011.
- [22] R. Li, J. Yu, S. Zhang, F. Bao, P. Wang, X. Huang, and J. Li, "Bayesian network analysis reveals alterations to default mode network connectivity in individuals at risk for alzheimer's disease," *PLoS One*, vol. 8, no. 12, p. e82104, 2013.
- [23] R. Li, X. Wu, K. Chen, A. Fleisher, E. Reiman, and L. Yao, "Alterations of directional connectivity among resting-state networks in alzheimer disease," *Am J Neuroradiol*, 2012.
- [24] X. Li, D. Coyle, L. Maguire, D. Watson, and T. McGinnity, "Gray matter concentration and effective connectivity changes in alzheimers disease: A longitudinal structural mri study," *Neuroradiology*, vol. 53, no. 10, pp. 733–748, 2011.
- [25] L. Zhou, L. Wang, L. Liu, P. Ogunbona, and D. Shen, "Discriminative brain effective connectivity analysis for alzheimers disease: A kernel learning approach upon sparse gaussian bayesian network," in *CVPR*, 2013.
- [26] ——, "Max-margin based learning for discriminative bayesian network from neuroimaging data," in *MICCAI*, 2014.
- [27] J. Kim, W. Zhu, L. Chang, P. Bentler, and T. Ernst, "Unified structural equation modeling approach for the analysis of multisubject, multivariate functional mri data," *Human Brain Mapping*, vol. 28, pp. 85–93, 2007.
- [28] K. Friston, L. Harrison, and W. Penny, "Dynamic causal modeling," *Neuroimage*, vol. 19, pp. 1273–1302, 2003.
- [29] A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic, "Free energy score spaces: Using generative information in discriminative classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1249–1262, 2012.
- [30] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, 2007.
- [31] J. Krapac, J. Verbeek, and F. Jurie, "Modeling spatial layout with fisher vectors for image categorization," in *ICCV*, 2011.
- [32] V. Sydorov, M. Sakurada, and C. Lampert, "Deep fisher kernels - end to end learning of the fisher kernel gmm parameters," in *CVPR*, 2014.
- [33] L. Maaten, "Learning discriminative fisher kernels," in *ICML*, 2011, pp. 217–224.
- [34] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 131–159, 2002.
- [35] L. Wang, "Feature selection with kernel class separability," *IEEE TPAMI*, vol. 30, no. 9, pp. 1534–1546, 2008.
- [36] X. Liu, L. Wang, J. Yin, E. Zhu, and J. Zhang, "An efficient approach to integrating radius information into multiple kernel learning," *IEEE TSMC-B*, 2012.
- [37] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2007.
- [38] R. Peharz and F. Pernkopf, "Exact maximum margin structure learning of bayesian networks," in *ICML*, 2012.
- [39] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.
- [40] A. Shojaei and G. Michailidis, "Penalized likelihood methods for estimation of sparse high dimensional directed acyclic graphs," *Biometrika*, vol. 97, no. 3, pp. 519–538, 2010.
- [41] F. Fu and Q. Zhou, "Learning sparse causal gaussian networks with experimental intervention: Regularization and coordinate descent," *Journal of the American Statistical Association*, vol. 108, no. 501, pp. 288–300, 2013.
- [42] ADNI, <http://www.adni-info.org>.
- [43] N. Kabani, J. MacDonald, C. Holmes, and A. Evans, "A 3d atlas of the human brain," *Neuroimage*, vol. 7, pp. S7–S17, 1998.
- [44] B. Tijms, P. Seris, D. Willshaw, and S. Lawrie, "Similarity-based extraction of individual networks from gray matter mri scans," *Cereb Cortex*, vol. 22, no. 7, pp. 1530–1541, 2012.
- [45] <Http://www.cs.huji.ac.il/site/labs/compbio/Repository/>.
- [46] D. Mackey, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- [47] D. Margaritis and S. Thrun, "Bayesian network induction via local neighborhoods," in *Proceedings of NIPS*, 1999.
- [48] I. Tsamardinos and C. Aliferis, "Towards principled feature selection: Relevancy, filters and wrappers," in *International Workshop on Artificial Intelligence and Statistics*, 2003.
- [49] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. Springer-Verlag, 1993.
- [50] K. Supekar, V. Menon, D. Rubin, M. Musen, and M. Greicius, "Network analysis of intrinsic functional brain connectivity in alzheimer's disease," *PLoS Computational Biology*, vol. 4, no. 6, pp. 1–11, 2008.
- [51] K. Wang, M. Liang, L. Wang, L. Tian, X. Zhang, K. Li, and T. Jiang, "Altered functional connectivity in early alzheimer's disease: a resting-state fmri study," *Human Brain Mapping*, vol. 28, no. 10, pp. 967–978, 2007.
- [52] R. Gould, B. Arroyo, R. Brown, A. Owen, E. Bullmore, and R. Howard, "Brain mechanisms of successful compensation during learning in alzheimer disease," *Neurology*, vol. 67, no. 6, pp. 1011–1017, 2006.

## APPENDIX

**Proposition 1.** Given a sparse Gaussian Bayesian Network parameterized by  $\Theta$  and its associated directed graph  $\mathcal{G}$  with  $m$  nodes, the graph  $\mathcal{G}$  is DAG if and only if there exist some  $o_i$  ( $i = 1, \dots, m$ ) and  $\Upsilon \in \mathbb{R}^{m \times m}$ , such that for arbitrary  $\Delta > 0$ , the following constraints are satisfied:

$$o_j - o_i \geq \frac{\Delta}{m} - \Upsilon_{ij}, \quad \forall i, j \in \{1, \dots, m\}, \quad i \neq j \quad (1a)$$

$$\Upsilon_{ij} \geq 0, \quad (1b)$$

$$\Upsilon_{ij} \times \Theta_{ij} = 0, \quad (1c)$$

$$\Delta \geq o_i \geq 0. \quad (1d)$$

**Proof.** As is known, a Bayesian network is equivalent to a topological ordering (Chapter 8, Section 8.1 on Page 362 in [37]). Therefore, we prove Proposition 1 by showing that i) Eqn. (1a ~ 1d) lead to a topological ordering (the necessary condition), and ii) a topological ordering from a DAG can meet the requirements in Eqn. (1a ~ 1d) (sufficient condition).

First, we prove the necessary condition by contradiction (Fig. 6). We consider three cases for two nodes  $j$  and  $i$ . Case 1) the nodes  $j$  and  $i$  are directly connected. If there is an edge from node  $i$  to node  $j$ , the parameter  $\Theta_{ij}$  is then non-zero, and thus  $\Upsilon_{ij}$  must be zero. According to Eqn. (1a), we have  $o_j > o_i$ . If at the same time, there is an edge from node  $j$  to node  $i$ , similarly we have  $o_i > o_j$ , which contradicts with  $o_j > o_i$ , and therefore is impossible. Case 2: the nodes  $j$  and  $i$  are not directly linked but connected by a path. Suppose there is a directed path  $P1$  from node  $i$  to node  $j$ , where  $P1$  is composed of nodes  $k_1, k_2, \dots, k_{m_1}$  in order. Following the above proof, we can have  $o_j > o_{k_{m_1}} > \dots > o_{k_1} > o_i$ . If at the same time another directed path  $P2$  links node  $j$  to node  $i$ , where  $P2$  is composed of nodes  $l_1, l_2, \dots, l_{m_2}$  in order, similarly we have  $o_i > o_{l_{m_2}} > \dots > o_{l_1} > o_j$ , making the contradiction. Case 3) If there is no edge between node  $i$  and node  $j$ , by definition  $\Theta_{ij} = 0$ . It is straightforward to see Eqn. (1b) and Eqn. (1c) hold for any arbitrary non-negative  $\Upsilon_{ij}$ . Moreover, for any  $o_i$  and  $o_j$  satisfying Eqn. (1d), we can show that as long as  $\Upsilon_{ij} \geq (\frac{1}{m} + 1)\Delta$  (which is positive), Eqn. (1a) will always hold. This is further explained as follows. By Eqn. (1d), we have  $-\Delta \leq o_j - o_i \leq \Delta$ . For Eqn. (1a) to be always held, we need some  $\Upsilon_{ij}$  such that  $o_j - o_i \geq -\Delta \geq \frac{\Delta}{m} - \Upsilon_{ij}$ , which requires  $\Upsilon_{ij} \geq (\frac{1}{m} + 1)\Delta$ . Therefore, there exist a set of  $o_i$  and  $\Upsilon$  valid for Eqn. (1a ~ 1d) when no edge links node  $i$  and node  $j$ . In sum, Eqn. (1a ~ 1d) show a topological ordering, that is, if node  $j$  comes after node  $i$  (that is,  $o_j > o_i$ ) in the ordering, there can not be a link from node  $j$  to node  $i$ , which guarantees the acyclicity.

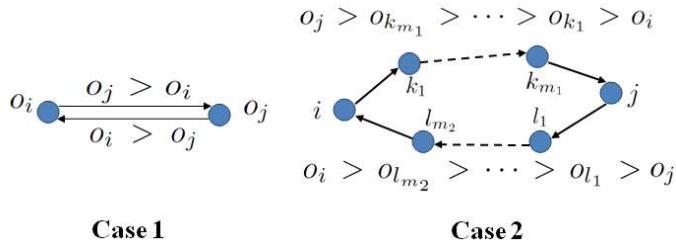


Fig. 6. Explanation of our ordering based DAG constraint.

Now let us consider the sufficient condition. if  $\mathcal{G}$  is a DAG, we can obtain some topological ordering  $(1, 2, \dots, m)$  from it. Let  $\tilde{o}_i$  be the index of node  $i$  in this ordering. Setting  $o_i = (\tilde{o}_i - 1)\frac{\Delta}{m}$  ( $\forall i \in \{1, \dots, m\}$ ), we have  $\min(o_i) = (1 - 1)\frac{\Delta}{m} = 0$  and  $\max(o_i) = (m - 1)\frac{\Delta}{m} \leq \Delta$ . If node  $j$  comes after node  $i$ , we have  $o_j - o_i \geq \frac{\Delta}{m} \geq \frac{\Delta}{m} - \Upsilon_{ij}$ . If node  $j$  comes before node  $i$ , we can always set  $\Upsilon_{ij}$  sufficiently large to satisfy Eqn. (1a ~ 1d). Therefore, from a DAG, we can always construct a set of ordering variables that satisfy Eqn. (1a ~ 1d).

Combining the proofs above, Eqn. (1a ~ 1d) are the sufficient and necessary condition for a directed graph  $\mathcal{G}$  to be DAG.  $\square$

**Proposition 2.** The optimization problem in Eqn. (2) (i.e., Eqn. (4.2) in the paper) is iteratively solved by alternate optimizations of (i)  $\mathbf{o}$  and  $\Upsilon$  with  $\Theta$  fixed, and (ii)  $\Theta$  with  $\mathbf{o}$  and  $\Upsilon$  fixed. This optimization converges and the output  $\Theta^*$  is DAG when  $\lambda_{dag} > \frac{2m(m-2)(n-1)^2 + m\lambda_1(2n-2-\lambda_1)}{\lambda_1(1+m)\Delta}$ , where  $m$  is the number of nodes and  $n$  is the number

of samples.

$$\begin{aligned} & \min_{\Theta, \mathbf{o}, \Upsilon} \sum_{i=1}^m \|\mathbf{x}_{:,i} - \mathbf{PA}_i^\top \boldsymbol{\theta}_i\|_2^2 + \lambda_1 \|\boldsymbol{\theta}_i\|_1 + \lambda_{dag} \boldsymbol{\epsilon}_i^\top |\boldsymbol{\theta}_i| \\ & \text{s.t. } o_j - o_i \geq \frac{\Delta}{m} - \Upsilon_{ij}, \forall i, j \in \{1, \dots, m\}, \quad i \neq j \\ & \quad 0 \leq o_i \leq \Delta, \quad \Upsilon_{ij} \geq 0 \end{aligned} \quad (2)$$

Here  $\mathbf{o}$  and  $\Upsilon$  are the variables defined in the DAG constraint in Section 4.2, and  $\Theta$  is the model parameters of SGBN. The vector  $\boldsymbol{\epsilon}_i$  denotes the  $i$ -th column of the matrix  $\Upsilon$ , and  $|\boldsymbol{\theta}_i|$  the component-wise absolute value of the  $i$ -th column of  $\Theta$ . Other parameters are defined in Table 1 in the paper.

**Proof.** In the following, we prove that:

- 1) The alternate optimization in Eqn. (2) converges.
- 2) The solution  $\Theta^*$  of Eqn. (2) is DAG when  $\lambda_{dag}$  is sufficiently large.

Let us denote  $f(\Theta, \mathbf{o}, \Upsilon) = \sum_{i=1}^m \|\mathbf{x}_{:,i} - \mathbf{PA}_i^\top \boldsymbol{\theta}_i\|_2^2 + \lambda_1 \|\boldsymbol{\theta}_i\|_1 + \lambda_{dag} \boldsymbol{\epsilon}_i^\top |\boldsymbol{\theta}_i|$ .

**First,** we prove Eqn. (2) converges by showing that (i)  $f(\Theta, \mathbf{o}, \Upsilon)$  is lower bounded; and (ii)  $f(\Theta^{(t+1)}, \mathbf{o}^{(t+1)}, \Upsilon^{(t+1)}) \leq f(\Theta^{(t)}, \mathbf{o}^{(t)}, \Upsilon^{(t)})$ , meaning that the function value will monotonically decrease with the iteration number  $t$ .

It is easy to see that  $f(\Theta, \mathbf{o}, \Upsilon)$  is lower bounded by 0, since each term in  $f(\Theta, \mathbf{o}, \Upsilon)$  is non-negative. And the second point can be proven as follows. At the  $t$ -th iteration, we update  $\Theta$  by

$$\begin{aligned} \Theta^{(t+1)} &= \arg \min_{\Theta} \sum_{i=1}^m \|\mathbf{x}_{:,i} - \mathbf{PA}_i^\top \boldsymbol{\theta}_i\|_2^2 + \lambda_1 \|\boldsymbol{\theta}_i\|_1 + \lambda_{dag} \boldsymbol{\epsilon}_i^{(t)\top} |\boldsymbol{\theta}_i| \\ &= \arg \min_{\Theta} f(\Theta, \mathbf{o}^{(t)}, \Upsilon^{(t)}). \end{aligned} \quad (3)$$

It holds that  $f(\Theta^{(t+1)}, \mathbf{o}^{(t)}, \Upsilon^{(t)}) \leq f(\Theta^{(t)}, \mathbf{o}^{(t)}, \Upsilon^{(t)})$ . Also it is noted that  $\Theta^{(t+1)}$  is an achievable global minimum of  $\Theta$  since  $f(\Theta, \mathbf{o}^{(t)}, \Upsilon^{(t)})$  is a convex function with respect to  $\Theta$ . Similarly, we then update  $\mathbf{o}$  and  $\Upsilon$  by

$$\begin{aligned} \{\mathbf{o}^{(t+1)}, \Upsilon^{(t+1)}\} &= \arg \min_{\mathbf{o}, \Upsilon} f(\Theta^{(t+1)}, \mathbf{o}, \Upsilon) \\ &\text{s.t. } o_j - o_i \geq \frac{\Delta}{m} - \Upsilon_{ij}, \forall i, j \in \{1, \dots, m\}, \quad i \neq j \\ &\quad 0 \leq o_i \leq \Delta, \quad \Upsilon_{ij} \geq 0. \end{aligned} \quad (4)$$

It holds that  $f(\Theta^{(t+1)}, \mathbf{o}^{(t+1)}, \Upsilon^{(t+1)}) \leq f(\Theta^{(t+1)}, \mathbf{o}^{(t)}, \Upsilon^{(t)})$ . Also,  $f(\Theta^{(t+1)}, \mathbf{o}, \Upsilon)$  is a linear function with respect to  $\mathbf{o}$  and  $\Upsilon$ . Consequently we have

$$f(\Theta^{(t+1)}, \mathbf{o}^{(t+1)}, \Upsilon^{(t+1)}) \leq f(\Theta^{(t+1)}, \mathbf{o}^{(t)}, \Upsilon^{(t)}) \leq f(\Theta^{(t)}, \mathbf{o}^{(t)}, \Upsilon^{(t)}).$$

Therefore, the optimization problem in Eqn. (2) is guaranteed to converge with the alternate optimization strategy, because the objective function is lower-bounded and monotonically decreases with the iteration numbers.

**Second,** we prove that when  $\lambda_{dag} > \frac{2m(m-2)(n-1)^2 + m\lambda_1(2n-2-\lambda_1)}{\lambda_1(1+m)\Delta}$ , the output  $\Theta^*$  is guaranteed to be DAG. This could be proven by contradiction. Suppose that such a  $\lambda_{dag}$  does not lead to a DAG, say,  $\Upsilon_{ji} \times \Theta_{ji}^* \neq 0$  for at least one pair of nodes  $i$  and  $j$ , with  $\Theta_{ji}^* \neq 0$  and  $\Upsilon_{ji} > 0$ . Without loss of generality, we assume  $\Upsilon_{ji} \geq (\frac{1}{m} + 1)\Delta$  (where  $\Delta$  is an arbitrary positive number), so the ordering constraints in Eqn. (2) always hold regardless of the variables  $o_i$  and  $o_j$ . This is because  $o_i$  and  $o_j$  are constrained by  $0 \leq o_i \leq \Delta$  and  $0 \leq o_j \leq \Delta$ , and  $o_j - o_i \geq -\Delta = \frac{1}{m}\Delta - (\frac{1}{m} + 1)\Delta$ . Based on the first-order optimality condition,  $\Theta_{ji}^* \neq 0$  i.f.f.  $2 \left| \left( \mathbf{x}_{:,i} - \mathbf{PA}_{i(\setminus j,:)}^\top \boldsymbol{\theta}_{i(\setminus j)}^* \right)^\top \mathbf{x}_{:,j} \right| - (\lambda_1 + \lambda_{dag} \Upsilon_{ji}) > 0$ . Here,  $\mathbf{PA}_{i(\setminus j,:)}$  denotes the elements in the matrix  $\mathbf{PA}_i$  with the  $j$ -th row removed (i.e., parents of the node  $i$  without the

node  $j$ ), and  $\boldsymbol{\theta}_{i \setminus j}^*$  denotes the elements in  $\boldsymbol{\theta}_i^*$  without  $\Theta_{ji}^*$ . However, it can be shown that,

$$\begin{aligned}
\left| \left( \mathbf{x}_{:,i} - \mathbf{PA}_{i(\setminus j,:)}^\top \boldsymbol{\theta}_{i \setminus j}^* \right)^\top \mathbf{x}_{:,j} \right| &\leq |\mathbf{x}_{:,i}^\top \mathbf{x}_{:,j}| + \left| \boldsymbol{\theta}_{i \setminus j}^{*\top} \mathbf{PA}_{i(\setminus j,:)} \mathbf{x}_{:,j} \right| \\
&= |\mathbf{x}_{:,i}^\top \mathbf{x}_{:,j}| + \sum_{k=1, k \neq i, j}^m |\Theta_{ki}^* \mathbf{x}_{:,k}^\top \mathbf{x}_{:,j}| \\
&\leq (n-1) + (m-2)(n-1) \max |\Theta_{ki}^*| \\
&\leq (n-1) + \frac{(m-2)(n-1)^2}{\lambda_1}.
\end{aligned} \tag{5}$$

The second last inequality holds due to the normalization of features  $\mathbf{x}_{:,i}$  (to zero mean and unit std). The last inequality holds because  $\max |\Theta_{ki}^*| \leq \|\boldsymbol{\theta}_i^*\|_1 \leq \frac{1}{\lambda_1} \left( \|\mathbf{x}_{:,i} - \mathbf{PA}_i^\top \boldsymbol{\theta}_i^*\|_2^2 + \lambda_1 \|\boldsymbol{\theta}_i^*\|_1 + \lambda_{dag} \boldsymbol{\epsilon}_i^{*\top} |\boldsymbol{\theta}_i^*| \right) = \frac{1}{\lambda_1} f(\boldsymbol{\Theta}^*, \mathbf{o}^*, \boldsymbol{\Upsilon}^*) \leq \frac{1}{\lambda_1} f(\mathbf{0}, \mathbf{o}^*, \boldsymbol{\Upsilon}^*) = \frac{1}{\lambda_1} \mathbf{x}_{:,i}^\top \mathbf{x}_{:,i} = \frac{n-1}{\lambda_1}$ . With the given  $\lambda_{dag}$ , Eqn. (5) results in

$$2 \left| \left( \mathbf{x}_{:,i} - \mathbf{PA}_{i(\setminus j,:)}^\top \boldsymbol{\theta}_{i \setminus j}^* \right)^\top \mathbf{x}_{:,j} \right| - (\lambda_1 + \lambda_{dag} \boldsymbol{\Upsilon}_{ij}) < 0,$$

which contradicts the above first-order optimality condition with  $\Theta_{ji}^* \neq 0$ . Therefore, when  $\lambda_{dag}$  is sufficiently large, the output  $\boldsymbol{\Theta}^*$  is guaranteed to be DAG.

Summing up the proofs above, the alternate optimization of Eqn. (2) converges and the output  $\boldsymbol{\Theta}^*$  is guaranteed to be DAG when  $\lambda_{dag}$  is sufficiently large.  $\square$